Universitat Ramon Llull

TESI DOCTORAL

Títol

OPTIMITZACIÓ PERCEPTIVA
DELS SISTEMES DE SÍNTESI DE LA PARLA
BASATS EN SELECCIÓ D'UNITATS MITJANÇANT
ALGORISMES GENÈTICS INTERACTIUS ACTIUS

Realitzada per Lluís Fomiga i Fanals

en el Centre Enginyeria i Arquitectura La Salle

i en el Departament Departament de Tecnologies Audiovisuals

Dirigida per Dr. Francesc Alías i Pujol

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es
www.url.es

“De vegades, la vida et colpeja amb un maó al cap. No perdin la fe. Estic convençut que l'únic que em va permetre seguir va ser que jo estimava el que feia. Han de trobar això que estimen. I això és tan vàlid pel treball com per l’amor. El seu treball omplirà gran part de les seves vides i l'única manera de sentir-se realment satisfet és fer allò que creuen que és un gran treball. I l'única forma de fer un gran treball és estimant el que fan. Si encara no l'han trobat, segueixin buscant. No es detinguin”

Steve Jobs – Discurs Honoris Causa Universitat de Stanford (12 de Juny de 2005)

“La vida està unida si es posa el cor en tot allò que es fa. El cor no com a sentiment, sinó com a desig insuprimible de felicitat, de bé, de veritat, de justícia. Aquest desig que un sempre alberga i al que tu sol no pots donar una resposta plena. Poder posar en joc tot el cor-és a dir, el teu desig de felicitat al complet-, en tot el que fas: tant en les situacions fàcils com en les difícils, en el cansament o en la diversió, en la família o a la feina . El cor, com a desig irreductible de la veritat, de la bellesa, de ser estimats i d'estimar. [...] Tot això, però, no n'hi ha prou, perquè sols no resistim. Fins i tot aquell que actua amb les millors intencions és incapaç. Cal que aquest “allò més gran” sigui una experiència, sigui Algú present a qui respondre. No una cosa que penso o que sento. [...] És necessari no estar sols. Cal un punt de suport. Necessitem una pertinença. [...] Amb el temps, la satisfacció no se li nega a qui s'equivoca, sinó que se li nega a qui no té el sentit del misteri en la seva vida, és a dir, quelcom més gran que està present, que és un companyia a la qual pertànyer”

Enzo Piccinini

“Per entendre si una hipòtesi és veritable cal tenir la certesa que s'assolirà la solució, cal emprendre el camí amb seguretat, d'aquesta manera, si la hipòtesi és veritable trobaràs la meta. Si comences a dir: «No .. serà una il.lusió », encara que sigui veritable ja no la trobaràs. Davant d'una proposta, l'única manera d'arribar a entendre si és vertadera o no és prendre-la seriosament.”

Luigi Guissani – El Sentit Religiós (Ed. Encuentro, 2008)

A la meva dona, la Rosa Maria

Agraïments

Moltes coses han canviat a la meva vida professional des que vaig començar la meva relació amb la recerca amb una col·laboració departamental l'any 2003. Si quelcom he après durant aquests anys, és que la recerca és un art, i no una tècnica, que l'única cosa que guia de debò una recerca és una autèntica passió pel coneixement, i que aquest coneixement, dóna molt més fruits que qualsevol producte final que puguis desenvolupar amb la millor tècnica possible. Per això, vull escriure unes línies a aquelles persones que han despertat en mi una autèntica passió pel coneixement i han esdevingut imprescindibles per la consecució d'aquesta Tesi Doctoral.

Primer de tot vull agrair tots aquests anys de treball i aprenentatge a en Francesc Alías, el meu director de tesi. Francesc, al teu cantó, m'has exercit de mestre, amic, company i confident. M'has curtit en la recerca i a base de projectes, m'has ensenyat rigor i a tenir una actitud crítica, a tenir una mentalitat interdisciplinària i no encasellar-me. Aquesta tesi no seria una realitat si no m'haguessis permès col·laborar en la teva tesi però sobretot si no haguessis confiat en mi i en el que fèiem. Ara, m'ha arribat l'hora d'aixecar el vol i abandonar el niu, però sempre aquest grup de recerca serà la meva llar, i això és gràcies al teu esperit. Deixo enrere un gran mestre i investigador, però m'alegra saber que el que has fet per mi ho faràs per les noves generacions que venen al darrera. Gràcies!

El segon agraïment d'aquesta tesi doctoral és a en Xavier Llorà. Ell fou qui va penjar 10 anys enrere una oferta de treball final de carrera que em va captivar. Gràcies a aquella oferta vaig poder començar a treballar en un ampli ventall de disciplines teòriques i pràctiques. Xevi, no és fàcil mantenir una col·laboració de treball a 7300 km. de distància, i menys durant tant de temps. Gràcies pel teu coneixement, compromís, experiència i facilitat de poder treballar amb tu i resoldre qualsevol circumstància.

Si aquesta tesi ha pogut fer un cop de puny sobre la taula i sortir endavant, en l'aspecte

personal ho haig d'agrair a la padrina de la meva filla i al seu marit, l'Anna Garriga i en Jorge Martínez. Gràcies a la vostra companyia m'heu fet veure que recerca, docència i benefici econòmic no són paraules excloents a la universitat. És més, la integració d'aquests conceptes és el que defineix la universitat com a nucli dur d'unitat en el coneixement. Gràcies per ensenyar-me que ser universitari es més que una professió, sinó una manera de viure. Juntament amb vosaltres voldria agrair explícitament a tota la gent d'Universitas per la seva companyia aquests dos últims anys. És impossible mencionar-los a tots però almenys voldria agrair a Carmine di Martino el seu treball "El coneixement és un esdeveniment", a la Guadalupe Arbona per proposar-nos treballar els textos, en Javier Prades per la seva actitud crítica quan ens perdíem en tonteries, a en Javier Ortega per la preocupació mostrada aquests últims mesos difícils i a en Pepe i la Manoli per obrir-nos casa seva i la seva família. En aquesta línia voldria també agrair l'acolliment de l'Enrico Magistretti i la Graziella, per ensenyar-nos una nova manera de viure basada en l'acolliment.

De manera aïllada i sincera voldria agrair a en José Miguel García ser el meu confident i conseller en tot moment. José Miguel: Te debo mucho, y se que la única manera de devolverte lo que me has dado es dar el testimonio de todo lo que me has enseñado a mis amistades, compañeros, possibles alumnos e hijos. Muchas gracias!

També voldria tenir unes paraules pel CLU de la Salle i els seus amics, gràcies per haver-me ajudat a trencar el dualisme entre saber i creure a la universitat. Gràcies Pepe Albín, Luís Albín, Borja Baeza, Ignasi Viñas i la Sílvia Aybar entre d'altres.

De tota la gent amb qui he pogut treballar aquests anys, voldria fer un agraïment molt especial i sincer a en Jordi Adell. El seu coneixement, experiència i esperit crític han significat un impuls molt important per tirar aquesta tesi endavant i poder detectar els punts dèbils abans que ningú. Jordi, valoro molt la teva ajuda desinteressada i espero poder-te la tornar en un futur no molt llunyà.

També voldria agrair la seva amistat i suport a l'Àlex Trilla aquest últim any i mig. Àlex, la teva amistat m'ha permès posar un mirall a davant per veure i recordar qui era i per a què feia les coses. Gràcies per ocupar-te i responsabilitzar-te de més coses de les que et tocaven quan jo no podia amb tot. I gràcies pel teu bon humor i esperança. Sobretot, segueix amb aquest esperit crític d'ara endavant.

Nogensmenys, voldria tenir paraules agraïments pels companys de vaixell en l'inici d'aquest viatge: Carlos Monzo i Xavier Gonzalvo i pels altres companys del grup de recerca que han col·laborat en la recerca ni que sigui fent proves perceptives: Borja Martínez, Àngel Calzada, Jordi Massana, Ignasi Iriondo, Joan Claudi Socoró, Oriol Guasch, Carles Vilella, Pere Artís, Lluís Cortés, Marc Arnela, Marc Rovira i Gabriel Fernández entre

d'altres. Especialment voldria agrair a en Xavier Sevillano el seu suport i discussions en algorismes de *clustering*, a en Santi Planet per haver desenvolupat la plataforma TRUE i a l'Elisa Martínez i en David Badía per haver confiat el projecte SALERO a les capacitats de tres estudiants de doctorat amb una altíssima motivació.

Voldria fer un agraïment explícit a la institució "Enginyeria La Salle" i la Universitat Ramon Llull per haver-me acollit durant aquests últims cinc anys i mig com a treballador i investigador i alhora haver-me proporcionat tot el material (corpus, equips de simulació, cursos i formació) que necessitava. No sé si els nostres camins es tornaran a creuar, però si és així segur que serà una gran notícia per mi. També voldria fer esment a la Universitat Politècnica de Catalunya i a la Universitat de Vigo per haver cedit els seus corpus a les competicions *Albayzín* en les dues últimes edicions. El seu gest ha permès poder confrontar la meva tesi doctoral fora de la Salle per poder analitzar la seva viabilitat o aplicabilitat fora del grup de recerca.

A nivell personal voldria agrair a la gent qua ha passat per la Masia/La Trama la seva companyia i paciència durant aquesta travessia: L'Eva Aguilera, en Ferran Riera, la Teresa Fernández, en Jordi Cabanes, la Glòria Arnau, en Paco Barquero, la Meritxell Mas, en Joan Lluís Pijoan, la Marta Martínez i a l'Enric Cantín i la Mònica Forroy.

Per anar acabant voldria agrair a totes aquelles persones que m'han fet tant els ajustos com les proves de validació perceptives. A part de la gent del grup que ja he mencionat, voldria tenir paraules d'agraïment per aquelles persones de fora del grup: German Cobo, Maria Alsina, Esther Adelantado, Matthias Eibel, José Antonio Montero, Francesc Escudero, David Garcia, Berta Martínez, Ester Cierco, Llúcia Sanz, Gemma Roig, Pau Bergadà, Pere Ordis, Valentí Alonso, Janine Sprünker, Tomás Schwab, Glòria Arnau, Adrià Casajús, David Maya, Josep Maria Ribes, Jordi Turon, Gemma Pertegaz, Lluís Bou, Josep Maria Illa, Ingrid Fàbrega, Ivan Traus, Lídia Serenó i Ricard Aquilué. Sense la vostra ajuda no hagués pogut acabar la tesi.

Abans d'acabar voldria tenir unes paraules d'agraïment a la meva família per la seva paciència durant aquests anys: Als meus pares Joaquim i Maria Dolors, als meus avis Lluís, Enric, Carme i Carme, a la Marianna, en Carlos, la Maria, la Núria i el nou que vindrà. També a la meva família política: la Carme i la Montse, l'àvia Maria i la iaia Florentina, a en Mateu i de manera molt especial a la meva sogra la Maria Rosa: Gràcies per haver-nos fet la vida extraordinàriament fàcil aquests últims mesos encarregant-te de la casa i de la Tina.

En aquests moments també tinc molt present a la meva filla, la Florentina. Tot i que només tens vuit mesos en el moment que estic escrivint aquestes línies et vull agrair el fet

que siguis aquí, els teus somriures, el teu afecte, la teva estima i la il·lusió de futur que has posat a la meva vida. Espero ser un bon pare que et sàpiga correspondre.

En últim lloc vull agrair Tot Plegat a la meva dona, la Rosa Maria, a la qual dedico aquesta Tesi. Rosa, tu m'has ensenyat a Creure, a tenir Esperança, a Donar sense esperar res a canvi, a no deixar-me vèncer per la por, a no rendir-me en els moments difícils, a jutjar les coses des d'un punt de vista realista i apartar els somnis que no condueixen a res. A oblidar-me de dogmatismes i prejudicis i ser fidel a la Veritat malgrat escollir a vegades la opció més difícil. Junts hem forjat una manera de viure que no és corrent avui en dia. Aquesta Veritat ens ha alliberat, sense ser esclaus de res, i ens ha permès de viure d'una manera molt més Humana. Junts, hem descobert una sèrie de valors i principis, però el més important, hem descobert el Perquè o Per Qui d'aquests valors i principis, i només demano que siguem capaços de transmetre-ho a les noves generacions i especialment a la nostra filla, la Florentina. T'Estimo!

Girona a 14 de Març del 2011

Resum

Els sistemes de conversió de text en parla (CTP-SU) s'encarreguen de produir veu sintètica a partir d'un text d'entrada. Els CTP basats en selecció d'unitats (CTP-SU) recuperen la millor seqüència d'unitats de veu enregistrades prèviament en una base de dades (corpus). La recuperació es realitza mitjançant algorismes de programació dinàmica i una funció de cost ponderada (Hunt i Black, 1996). La ponderació de la funció de cost es realitza típicament de forma manual per part d'un expert (Clark *et al.*, 2007; Schröder *et al.*, 2009). No obstant, l'ajust manual resulta costós des d'un punt de vista de coneixement prèvi, i imprecís en la seva execució (Strom i King, 2008). Per tal d'ajustar els pesos de la funció de cost, aquesta tesi parteix de la prova de viabilitat d'ajust perceptiu presentada per Alías (2006) que empra algorismes genètics interactius actius (*active interactive Genetic Algorithm* - aiGA - (Llorà *et al.*, 2005*b*)). Aquesta tesi doctoral investiga les diferents problemàtiques que es presenten en aplicar els aiGAs en l'ajust de pesos d'un CTP-SU en un context real de selecció d'unitats. Primerament la tesi realitza un estudi de l'estat de l'art en l'ajust de pesos. Tot seguit, repassa la idoneïtat de la computació evolutiva interactiva per realitzar l'ajust revisant amb profunditat el treball previ (Alías, 2006). Llavors es presenten i es validen les propostes de millora. Les quatre línies mestres que guien les contribucions d'aquesta tesi són: la precisió en l'ajust dels pesos, la robustesa dels pesos obtinguts, l'aplicabilitat de la metodologia per qualsevol funció de cost i el consens dels pesos obtinguts incorporant el criteri de diferents usuaris. En termes de precisió la tesi proposa realitzar l'ajust perceptiu per diferents tipus (clústers) d'unitats respectant les seves peculiaritats fonètiques i contextuals. En termes de robustesa la tesi incorpora diferents mètriques evolutives (indicadors) que avaluen aspectes com l'ambigüitat en la cerca, la convergència d'un usuari o el nivell de consens entre diferents usuaris. Posteriorment, per estudiar l'aplicabilitat de la metodologia proposada s'ajusten perceptivament diferents pesos que combinen informació lingüística i simbòlica. La última contribució d'aquesta tesi estudia l'idoneïtat dels models latents per modelar les preferències dels diferents usuaris

i obtenir una solució de consens. Paral·lelament, per fer el pas d'una prova de viabilitat a un entorn real de selecció d'unitats es treballa amb un corpus d'extensió mitjana (1.9h) etiquetat automàticament. La tesi permet concloure que l'aiGA a nivell de clúster és una metodologia altament competitiva respecte les altres tècniques d'ajust presents en l'estat de l'art.

Resumen

Los sistemas de conversión texto-habla (CTH-SU) se encargan de producir voz sintética a partir de un texto de entrada. Los CTH basados en selección de unidades (CTH-SU) recuperan la mejor secuencia de unidades de voz grabadas previamente en una base de datos (corpus). La recuperación se realitza mediante algoritmos de programación dinámica y una función de coste ponderada (Hunt i Black, 1996). La ponderación de la función de coste se realiza típicamente de forma manual por parte de un experto (Clark *et al.*, 2007; Schröder *et al.*, 2009). Sin embargo, el ajuste manual resulta costoso desde un punto de vista de conocimiento previo e impreciso en su ejecución (Strom i King, 2008). Para ajustar los pesos de la función de coste, esta tesis parte de la prueba de viabilidad de ajuste perceptivo presentada por Alías (2006) que emplea algoritmos genéticos interactivos activos (*active interactive Genetic Algorithm* - aiGA- (Llorà *et al.*, 2005*b*)). Esta tesis doctoral investiga las diferentes problemáticas que se presentan al aplicar los aiGAs en el ajuste de pesos de un CTH-SU en un contexto real de selección de unidades. Primeramente la tesis realiza un estudio del estado del arte en el ajuste de pesos, posteriormente repasa la idoneidad de la computación evolutiva interactiva para realizar el ajuste revisando en profundidad el trabajo previo (Alías, 2006). Entonces se presentan y se validan las propuestas de mejora. Las cuatro líneas maestras que guían las contribuciones de esta tesis son: la precisión en el ajuste de los pesos, la robustez de los pesos obtenidos, la aplicabilidad de la metodología para cualquier función de coste y el consenso de los pesos obtenidos incorporando el criterio de diferentes usuarios. En términos de precisión la tesis propone realizar el ajuste perceptivo por diferentes tipos (*clusters*) de unidades respetando sus peculiaridades fonéticas y contextuales. En términos de robustez la tesis incorpora diferentes métricas evolutivas (indicadores) que evalúan aspectos como la ambigüedad en la búsqueda, la convergencia de un usuario o el nivel de consenso entre diferentes usuarios. Posteriormente, para estudiar la aplicabilidad de la metodología propuesta se ajustan perceptivamente diferentes pesos que combinan información lingüística y simbólica. La última contribución de esta tesis es-

tudia la idoneidad de los modelos latentes para modelar las preferencias de los diferentes usuarios y obtener una solución de consenso. Paralelamente, para dar el paso de una prueba de viabilidad a un entorno real de selección de unidades se trabaja con un corpus de extensión media (1.9h) etiquetado automáticamente. La tesis permite concluir que el aiGA a nivel de cluster es una metodología altamente competitiva respecto a las otras técnicas de ajuste presentes en el estado del arte.

Abstract

Text-to-Speech systems (TTS) produce synthetic speech from an input text. Unit Selection TTS (US-TTS) systems are based on the retrieval of the best sequence of recorded speech units previously recorded into a database (corpus). The retrieval is done by means of dynamic programming algorithm and a weighted cost function (Hunt i Black, 1996). An expert typically performs the weighting of the cost function by hand (Clark *et al.*, 2007; Schröder *et al.*, 2009). However, hand tuning is costly from a standpoint of previous training and inaccurate in terms of methodology (Strom i King, 2008). In order to properly tune the weights of the cost function, this thesis continues the perceptual tuning proposal submitted by Alías (2006) which uses active interactive Genetic Algorithms (aiGAs) (Llorà *et al.*, 2005*b*). This thesis conducts an investigation to the various problems that arise in applying aiGAs to the weight tuning of the cost function. Firstly, the thesis makes a deep revision to the state-of-the-art in weight tuning. Afterwards, the thesis outlines the suitability of Interactive Evolutionary Computation (IEC) to perform the weight tuning making a thorough review of previous work (Alías, 2006). Then, the proposals of improvement are presented. The four major guidelines pursued by this thesis are: accuracy in adjusting the weights, robustness of the weights obtained, the applicability of the methodology to any subcost distance and the consensus of weights obtained by different users. In terms of precision cluster-level perceptual tuning is proposed in order to obtain weights for different types (clusters) of units considering their phonetic and contextual properties. In terms of robustness of the evolutionary process, the thesis presents different metrics (indicators) to assess aspects such as the ambiguity within the evolutionary search, the convergence of one user or the level of consensus among different users. Subsequently, to study the applicability of the proposed methodology different weights are perceptually tuned combining linguistic and symbolic information. The last contribution of this thesis examines the suitability of latent models for modeling the preferences of different users and obtains a consensus solution. In addition, the experimentation is carried out through a medium si-

ze corpus (1.9h) automatically labelled in order fill the gap between the proof-of-principle and a real unit selection scenario. The thesis concludes that aiGAs are highly competitive in comparison to other weight tuning techniques from the state-of-the-art.

Glossari

aiGA *active interactive Genetic Algorithm*
ASF *Acoustic Space Formultation*
CART *Classification and Regression Tree*
cGA *compact Genetic Algorithm*
CMOS *Comparison Mean Opinion Score*
cPBIL *continuous Population Based Incremental Learning*
CTP Sistema de Conversió de Text en Parla
CTP-SU Sistema de Conversió de Text en Parla basat en Selecció d'Unitats
DTW *Dynamic Time Warping*
DUR Durada
EA *Evolutionary Algorithm*
EC *Evolutionary Computation*
EDA *Estimation of Distribution Algorithms*
ENE Energia
EM *Expectation Maximization*
GA *Genetic Algorithm*
GMM *Gaussian Mixtures Model*
GNE *Glottal-to noise excitation ratio*
GR *Gain Ratio*
GTM *Generative Topographinc Mapping*
HMM *Hidden Markov Models*
HNM *Harmonic plus Noise Model*
HUX *Half Uniform Crossover*
IEC *Interactive Evolutionary Computation*
IFF *Independent Feature Formulation*

IG *Information Gain*
iGA *interactive Genetic Algorithm*
IPA *International Phonetic Alphabet*
LPC *Linear Prediction Coding*
LS *Least Squares* - Mínims quadrats
LSF *Linear Spectral Frequencies*
LSP *Linear Spectral Pairs*
MAP *Maximum a posteriori*
MBROLA *Multi-Band Resynthesis Overlap and Add*
MFCC *Mel-Frequency Cepstral Coefficients*
ML *Maximum Likelihood*
MLR *Multilinear Regression*
MOS *Mean Opinion Score*
NNLS *Non-negative Least Squares*
PBIL *Problem Based Incremental Learning*
pdf *Probability Density Function*
PDS Processament digital del senyal
PIT *Pitch*
PLN Processament del llenguatge natural
POS *Part-of-Speech* - Categoria gramatical
PosInEG *Position in Entonation Group*
PosInSyll *Position in Syllable*
PosInWord *Position in Word*
PSOLA *Pitch Synchronous Overlap and Add*
qqplot Comparació quartil-quartil[1]
R^2 Coeficient de determinació
RMS *Root Mean Squared*
RMSE *Root Mean Squared Error*
RWS *Roulette Wheel Selection*
SAMPA *Speech Assessment Methods Phonetic Alphabet*
SGML *Standard Generalized Markup Language*
SOM *Self-Organizing Maps*
SVM *Support Vector Machine*
TDPSOLA *Time Domain Pitch Synchronous OverLap and Add*
ToBI *Tonal and Break Indices*

[1] En aquesta tesi doctoral, totes les comparacions quartil-quartil són respecte la distribució normal

TTP *Text-to-Phone*
TSVQ *Tree-Structured Vector Quantization*
UX *Uniform Crossover*
VQ *Vector Quantization*
VoQ *Voice Quality* - Qualitat de Veu
WFST *Weighted Finite State Transducers*
WSS *Weight Space Search*
WTISS *Weight Tuning Interface for Speech Synthesis*

Índex

Índex de taules

Índex de figures

Índex d'algorismes

CAPÍTOL 1

Introducció

1.1 Marc de treball general

El processament computacional de la parla i el llenguatge (*Speech and Language Processing*) és una disciplina que ha conviscut amb l'expansió de l'informàtica des dels seus orígens, immediatament posteriors a la II Guerra Mundial (1940-1950) (Jurafsky *et al.*, 2000). Aquesta disciplina contempla diferents aspectes tals com el processament del llenguatge natural (PLN), la lingüística computacional, el reconeixement i la síntesi de la parla, la identificació biomètrica de locutor, etc. Durant els transcurs dels anys, els resultats d'aquesta disciplina han passat de ser prototips de laboratori o curiositats en la ciència ficció a conviure quotidianament amb nosaltres (Holmes i Holmes, 2001). En termes generals, són diversos els exemples on les tecnologies basades en el processament de la parla i el llenguatge resulten habituals en l'actualitat: centres telefònics d'atenció al client (companyies telefòniques, administració pública, reserves *on-line*,etc.), sistemes de navegació i ajuda a la conducció (GPS, mans lliures, control de la ràdio i altres elements del cotxe,etc.), transcripció de discursos d'interès públic (política, notícies,etc.), traducció automàtica entre idiomes (Bonafonte *et al.*, 2006), personatges virtuals en l'àmbit de la informació i entreteniment (Haas i Thallinger, 2005; Alías *et al.*, 2005) o la mineria de textos (*text mining*) (Feldman i Sanger, 2008) per proporcionar anuncis adequats a l'usuari segons l'anàlisi automàtic d'opinions o xarxes socials (Jurafsky *et al.*, 2000).

Dins d'aquest escenari, els sistemes de conversió de text en parla (CTP) han jugat un rol fonamental en la interacció amb l'usuari (Taylor, 2009). En els últims set anys, i a partir de l'experiència del grup de recerca, s'ha participat en projectes de recerca i de transferència de tecnologia d'àmbits molt diversos, que com a factor comú, necessitaven d'un CTP. Aquests projectes cobrien un ampli ventall de d'aplicacions tals com la generació automàtica de continguts audiovisuals (SAM – CIDEM-RDITSCON04-0005,Alías *et al.* (2005)), SALERO – IST-FP6-027122 Haas i Thallinger (2005)), CuentaCuentos – TSI-070100-2008-19, GTM (2008*a*))), assistència a persones amb necessitats especials (Integra-TV 4all – FIT-350301-2004-3, Ceccaroni *et al.* (2005); INREDIS – CEN-20072011,Technosite (2009)) o la multimodalitat d'interacció en entorns col·laboratius (Reune-T – PPT-430000-2008-36, GTM (2008*b*)).

1.2 Marc de treball específic

Els sistemes de conversió de text en parla basats en selecció d'unitats (CTP-SU) (Hunt i Black, 1996) han estat l'aproximació més emprada per assolir veu sintètica d'alta qualitat durant l'última dècada (Cox *et al.*, 2002). Els CTP-SU són una de les tècniques amb més acceptació per la generació de veu sintètica juntament amb la tècnica de síntesi basada en models ocults de Markov (*Hidden Markov Models* - HMM) (Tokuda *et al.*, 2003). Les dues tècniques s'engloben dins dels sistemes de síntesi de tercera generació (Taylor, 2009). A continuació es detallen molt breument els seus principis fonamentals.

Els sistemes CTP-SU es basen en l'enregistrament, etiquetatge i recuperació de veu natural (corpus), centralitzant el focus de la investigació més en l'aprenentatge estadístic que en el processament del senyal, concretament en la recuperació d'unitats acústiques. Com a resultat, aquestes tècniques assoleixen una bona nota en naturalitat, riquesa i expressivitat de la producció de la parla (Kaszczuk i Osowski, 2009). No obstant, els CTP-SU presenten un decreixement considerable de la qualitat quan es sintetitzen textos o locucions amb una naturalesa diferent de les enregistrades (Stylianou i Syrdal, 2001), provocant així una pèrdua de flexibilitat i aplicabilitat del seu propòsit més general: generar veu sintètica d'alta qualitat a partir d'un text d'entrada qualsevol.

Per aquest motiu, últimament la recerca s'ha centrat en una nova tècnica de síntesi que basa el seu funcionament en l'aprenentatge estadístic dels paràmetres espectrals que conformen la veu, mitjançant models ocults de Markov (Tokuda *et al.*, 2003). Aquesta síntesi es coneix com a síntesi markoviana o HMM. Tanmateix, malgrat l'augment de la flexibilitat que permet aquesta tècnica de síntesi, la naturalitat de la parla sintètica dels

sistemes basats en HMM encara no ha superat la dels CTP-SU (Oura *et al.*, 2009) si s'empren corpus grans ($\geq$ 5h). En canvi, s'ha observat que els sistemes híbrids HMM + CTP-SU aconsegueixen superar la qualitat que obtenen els dos mètodes per separat (Lu *et al.*, 2009). No obstant, els sistemes híbrids també requereixen una fase de selecció en el seu disseny.

La ponderació de subcostos a la fase de disseny d'un CTP-SU és una de les qüestions més importants en per assolir una bona qualitat sintètica (Meron i Hirose, 1999; Campbell i Black, 1997). No obstant això, la seva optimització presenta una dificultat elevada (Campbell i Black, 1997), que en aquest cas es converteix en l'eix central d'aquesta tesi. Aquesta tesi presenta un pas més en l'ajust perceptiu de la ponderació de subcostos en sistemes de conversió de text a parla basats en selecció d'unitats. Concretament, es parteix de la prova de viabilitat descrita a Alías (2006) que aplicava els últims avenços en computació evolutiva interactiva (Llorà *et al.*, 2005*b*) per realitzar aquest ajust. L'aproximació proposada, realitza l'ajust de pesos de manera perceptiva emprant algorismes genètics interactius actius (*active interactive Genetic Algorithm* - aiGA). Cal remarcar que amb anterioritat a aquesta tesi s'ha col·laborat de manera molt propera en el desenvolupament d'aquesta prova de viabilitat (Alías *et al.*, 2003, 2004, 2006*a*). La investigació realitzada en aquesta tesi és una investigació multidisciplinar que combina el processament de la parla (Rabiner i Schafer, 1978), la computació evolutiva (Holland, 1975) i la interacció home-màquina (Takagi, 2001) (també anomenada interacció home-ordinador). A continuació es detalla de quina manera interaccionen les diferents disciplines.

1.2.1 Sistemes de conversió de text en parla basats en selecció d'unitats

Els sistemes de CTP-SU són l'evolució natural dels primers sistemes de síntesi concatenativa: els CTP per difonemes (Moulines i Charpentier, 1990). El funcionament dels CTP-SU es basa en la recuperació de la *millor* seqüència d'unitats d'una base de dades (corpus) per tal d'obtenir veu sintètica de la màxima qualitat possible (Hunt i Black, 1996). La degradació de la qualitat sintètica sovint és causada per la presència de discontinuïtats d'entonació, de volum, espectrals o contextuals en la composició del senyal sintètic a partir dels fragments recuperats del corpus (Sagisaka, 1988; Stylianou i Syrdal, 2001). Clàssicament, la selecció d'unitats es defineix com una cerca dinàmica (programació dinàmica) de les unitats del corpus per escollir el millor conjunt d'unitats entre les diverses candidates possibles (Hunt i Black, 1996). A tal efecte, es dissenya una funció de cost que pondera les diferències (subcostos) entre una especificació ideal i les unitats candidates (cost d'unitat) així com la idoneïtat de concatenar dues unitats de la base de dades (cost de concatenació) (Hunt i Black, 1996).

A la literatura, existeixen diverses aproximacions que intenten assolir una funció de cost òptima en termes de la qualitat sintètica que se'n deriva (Meron i Hirose, 1999; Chu i Peng, 2001; Toda *et al.*, 2006; Clark *et al.*, 2007). L'objectiu principal d'aquesta tesi persegueix dissenyar una metodologia que permeti trobar una ponderació òptima de la funció de cost per realitzar la selecció d'unitats seguint els treballs previs descrits a (Alías, 2006).

Paral·lelament a la ponderació dels pesos, la tesi aprofundeix sobre la tipologia dels subcostos de la funció, fruit de la discussió (Taylor, 2009; Steiner *et al.*, 2010) sobre la idoneïtat que siguin bé de naturalesa lingüística (Chu i Peng, 2001; Clark *et al.*, 2007) o bé de naturalesa acústica (Park *et al.*, 2003; Lee *et al.*, 2003).

1.2.2 Importància de la intervenció humana en el disseny dels sistemes CTP-SU

Degut a la complexitat de l'ajust de pesos de la funció de cost de selecció de manera automàtica, típicament aquest ajust es realitza manualment per part d'un expert (Clark *et al.*, 2007; Schröder *et al.*, 2009). Les aproximacions basades en un ajust manual obtenen bons resultats avui en dia en termes de qualitat sintètica (Schröder *et al.*, 2009). Tanmateix, aquest ajust manual resulta costós des d'un punt de vista de formació prèvia de la persona que realitza l'ajust, ja que requereix d'un alt grau d'expertesa, i alhora és imprecís degut a que la seva execució esdevé bastant artística i poc robusta (Strom i King, 2008). A més, l'ajust manual presenta dos biaixos addicionals: a les poques frases que s'empren en l'ajust i el biaix de l'ajust degut a les preferències acústiques pròpies de la persona que realitza l'ajust. En altres paraules, l'ajust a mà dificulta la cooperació de diferents persones, amb expertesa o sense, per tal d'optimitzar el sistema de forma més general. Per millorar la robustesa de l'ajust, a la literatura s'han plantejat diferents tècniques que sistematitzen l'obtenció dels pesos (Meron i Hirose, 1999; Chu i Peng, 2001; Lee *et al.*, 2003; Park *et al.*, 2003; Colotte i Beaufort, 2005; Toda *et al.*, 2006; Bellegarda, 2009). Aquestes tècniques poden comptar amb intervenció humana o ser íntegrament automàtiques. Les tècniques automàtiques no obtenen una qualitat òptima de manera contrastada respecte l'ajust a mà (Taylor, 2009). Per una altra banda en la literatura es demostra que l'ajust manual pot abastar com a molt l'ajust de 10 pesos simultàniament (Taylor, 2009), quan alguns dels millors sistemes de CTP-SU en l'actualitat tenen prop de 50 paràmetres de selecció que cal ajustar (Kaszczuk i Osowski, 2009).

L'altre problema de l'ajust manual és la generalització del criteri (un sol criteri) respecte les diferents peculiaritats fonètiques o contextuals de la selecció d'unitats. Aquest

aspecte guarda relació amb el biaix cap a les poques frases emprades en el disseny de la funció de cost. Els mètodes d'ajust automàtic ajusten la funció de cost respectant les característiques de cada fonema (Campillo *et al.*, 2005). En canvi, l'ajust manual fa pràcticament impossible obtenir pesos específics respectant les característiques fonètiques de cada unitat, conseqüentment es realitza una generalització de la funció de cost que no considera la informació relativa sobre la bondat dels subcostos en funció de la seva especificitat.

Aquesta tesi proposa que l'aspecte clau per afrontar el problema de la ponderació dels pesos de la funció de cost passa per la definició d'una metodologia capaç de combinar qualsevol formulació dels subcostos de forma cooperativa i perceptiva, tot respectant l'especificitat acústica de les unitats que conformen el corpus. A més, aquesta metodologia facilita l'accés de diferents perfils d'usuaris al procés, facilitant que experts i no experts realitzin l'ajust conjuntament.

En aquest sentit, aquesta tesi es basa en el fet que un disseny òptim de la funció de cost de selecció, ha de considerar dos subobjectius: *i*) incorporar de forma robusta la participació humana en l'ajust i *ii*) dissenyar una metodologia de selecció d'unitats que respecti l'especificitat de cada unitat a l'hora de ponderar els diferents subcostos.

1.2.3 Idoneïtat de la computació evolutiva interactiva

Entre els mètodes d'optimització més coneguts (Beightler *et al.*, 1979), la computació evolutiva interactiva (*Interactive Evolutionary Computation* - IEC) destaca per la seva capacitat de fusionar esforços humans i computacionals, sobretot quan la solució al problema és alhora no lineal i subjectiva (Takagi, 2001). La IEC se sustenta sobre la teoria de la computació evolutiva clàssica (*Evolutionary Computation* - EC) (Holland, 1975), inspirada en la teoria de l'evolució de Darwin. En aquest sentit, la EC cerca les solucions a un problema a través d'un mecanisme de "*supervivència-dels-millors*" que evoluciona una població d'individus imitant els processos presents a la natura: recombinació, mutació i selecció natural. La EC s'ha emprat amb èxit en el problema de l'ajust de pesos en CTP-SU a través de l'ús d'algorismes genètics (GAs)(Alías i Llorà, 2003). Malgrat que en principi pot semblar que una aproximació estocàstica dota de poca robustesa matemàtica la solució final (Goldberg, 1989), són extensos els entorns i les aplicacions on la EC i la IEC han demostrat, no només la seva robustesa matemàtica (Michalewicz, 1992; Takagi, 2001), sinó la seva capacitat d'aportar solucions òptimes, inclòs l'àmbit del processament del senyal (Durant *et al.*, 2004). La IEC segueix el mateix esquema que la EC incorporant un agent humà per realitzar la selecció natural de les millors solucions. La primera investigació sobre l'ajust

de pesos mitjançant IEC es va presentar a Alías *et al.* (2003, 2004). Concretament es va emprar un algorisme genètic interactiu (iGA). No obstant això, la naturalesa interactiva de l'iGA comportava problemes de fiabilitat en les solucions finals (Llorà *et al.*, 2005*b*). Aquests problemes generalment eren deguts a inconsistències provocades per la fatiga i la frustració en l'usuari degut al llarg procés evolutiu necessari per realitzar l'ajust o degut a repeticions en les avaluacions. Aquests problemes de fiabilitat provocaven que l'iGA fos una metodologia poc eficient. En un intent de millorar l'iGA, a Llorà *et al.* (2005*b*) es proposa l'iGA *actiu* o aiGA, una millora de l'iGA que adopta el concepte d'aprenentatge actiu (Cohn *et al.*, 1994) amb l'objectiu de reduir la fatiga i l'ambigüitat en l'evolució perceptiva d'un problema. L'aiGA, per a aconseguir-ho, modela l'espai perceptiu de l'usuari segons un graf de preferències. A partir d'aquest graf es pot obtenir un ordre complet induït de les solucions, generant així una funció d'avaluació (funció de *fitness*) artificial, capaç d'estimar la qualitat d'una solució sense la participació explícita de l'usuari. Aquest *fitness* artificial (*fitness* sintètic), permet reduir el nombre d'avaluacions repetitives així com escurçar la durada de les proves. Així, la combinació del *fitness* sintètic amb la interacció de l'usuari defineix un mètode eficient i robust d'optimització perceptiva d'un problema no lineal (Llorà *et al.*, 2005*b*). A Alías *et al.* (2006*a*) s'avalua la idoneïtat de l'aiGA per a afrontar el problema de l'ajust de pesos en CTP-SU. En aquell treball, també es va introduir una millora de la definició del propi aiGA pel fet d'identificar contradiccions de l'usuari (p.ex. $A > B$, $B > C$, $C > A$) com a cicles en el model de graf. Aquests cicles reflecteixen una degradació de la consistència del model, i per tant, de la qualitat de les solucions obtingudes. És en aquell treball on l'autor d'aquesta tesi doctoral va realitzar les seves primeres contribucions al problema plantejat. Concretament aquestes contribucions són la pròpia adaptació de l'aiGA al problema de l'ajust de pesos per CTP-SU i la detecció i eliminació de cicles en els grafs que modelen les avaluacions dels usuaris (Alías, 2006). No obstant, aquell primer treball es va limitar a ser una prova de viabilitat (*proof-of-principle*) deixant oberts diferents aspectes en la seva aplicabilitat en un entorn real de selecció d'unitats. Aquesta tesi doctoral identifica les limitacions del mètode oferint una proposta de millora.

1.3 Motivacions i objectius: precisió, robustesa, aplicabilitat i consens

La tècnica d'ajust de pesos de la funció de cost d'un CTP-SU basada en aiGA necessita d'una investigació més exhaustiva. Amb anterioritat a aquesta tesi, la investigació s'ha cenyit a l'avaluació de la viabilitat de la proposta a través d'un prototip dins d'un entorn

controlat (Alías, 2006). La selecció d'unitats es realitzava en un corpus de mida molt reduïda (8 min.) etiquetat a mà (supervisió manual) i per una tipologia concreta de subcostos (acústics).

L'objectiu principal d'aquesta tesi és dissenyar una metodologia d'optimització perceptiva de la funció de cost que satisfaci els principis de precisió, robustesa, consens i aplicabilitat a un entorn real de CTP-SU. En aquest sentit, aquesta tesi presenta diferents contribucions per satisfer cadascun d'aquests principis. Un cop obtinguda la nova metodologia s'estudiarà la seva idoneïtat en un entorn real de selecció d'unitats fent-la competir amb altres tècniques d'ajust vigents avui en dia, observant fins a quin punt resulta competitiva i aplicable en termes de la qualitat sintètica obtinguda

A continuació es descriuen els quatre aspectes que investiga aquesta tesi per tal de completar i millorar el primer prototip d'aplicació de la metodologia d'ajust de pesos basada en aiGA descrita a Alías (2006).

1.3.1 Precisió

Amb anterioritat a aquesta tesi doctoral, l'aiGA considera la generalització de la ponderació dels subcostos independentment de les diferents peculiaritats fonètiques o contextuals on s'ha de realitzar la selecció d'unitats (Alías, 2006). En canvi, en les metodologies d'ajust de pesos que no compten amb intervenció humana, es considera que la unitat marca certes especificats a l'hora de recuperar les unitats més adequades per a generar la veu sintètica (p.ex. la importància relativa de la durada en les oclusives). Aquestes especificats impliquen que una ponderació adequada per un tipus d'unitat, no ho sigui per una altra (p.ex. oclusiva vs. lateral) (Campillo *et al.*, 2005). En altres paraules, que cada unitat o tipus d'unitat hauria de tenir el seu vector de pesos associat. Tanmateix, en l'ajust perceptiu de la funció de cost, resulta pràcticament impossible obtenir un vector de pesos específic per a cada unitat, i en conseqüència, en general, es prioritza l'obtenció d'un vector de pesos global amb bons pesos perceptius, que diversos vectors de pesos d'alta precisió però baixa qualitat al no disposar d'un bon model automàtic d'avaluació de la qualitat. D'aquest fet sorgeix la primera motivació d'aquesta tesi: establir una metodologia perceptiva que permeti obtenir bons pesos perceptius en funció de les diferents especificats de la la funció de cost.

1.3.2 Robustesa

Robustesa dels pesos automàtics

D'entre tots els mètodes d'ajust de pesos automàtics de l'estat de l'art, el més conegut és l'ajust de pesos mitjançant una regressió lineal entre els diferents subcostos i les distàncies cepstrals (Meron i Hirose, 1999). Les metodologies d'ajust automàtic, però, obtenen uns pesos poc fiables segons les pròpies mètriques de mesura del model (R^2 i RMSE (Netter *et al.*, 1990)). Aquesta manca de robustesa en el modelat pot venir donada per dos motius: *i*) una normalització de les dades inadequada fa empitjorar els models de regressió lineal (Chatterjee i Hadi, 2006), i *ii*) no respectar el comportament diferencial dels subcostos en funció del context afegeix soroll en l'optimització, tal com s'ha comentat anteriorment en l'objectiu de precisió.

Per superar les restriccions lineals del model, Alías i Llorà (2003) proposen realitzar l'ajust automàtic mitjançant algorismes genètics, ja que són robustos davant problemes d'optimització en entorns sorollosos. Malgrat que el seu treball proporciona una millora notable en seleccionar unitats amb menor subcost, la proposta no proporciona cap índex de fiabilitat dels pesos obtinguts que mesuri el nivell de soroll en l'evolució ni tampoc s'avalua perceptivament la qualitat sintètica obtinguda, aspecte que s'estudia en aquesta tesi doctoral.

Respecte a la normalització dels subcostos, a la literatura no s'ha trobat definida amb detall quina és la millor normalització dels diferents subcostos per a la optimització lineal. En aquest sentit, aquesta tesi doctoral estudia diferents transformacions (Tukey, 1957) dins la funció de cost per normalitzar els diferents subcostos prosòdics i espectrals.

Robustesa de les solucions de l'aiGA

La prova de viabilitat de l'aiGA per l'ajust de pesos realitzada per Alías (2006) denota que les contradiccions de l'usuari proporcionen solucions poc consistents. Per tant, en aquell treball també s'indica que avaluar la fiabilitat de les solucions proporcionades pels usuaris en funció de la seva consistència aporta informació per obtenir bones solucions. No obstant això, reduir l'anàlisi a només la identificació de contradiccions dels usuaris pot resultar insuficient: un usuari pot no contradir-se i ser consistent durant el transcurs de la prova, i tanmateix, la seva actitud pot ser excessivament conservadora o dubitativa proporcionant un nombre elevat d'empats en les comparatives subjectives respecte la resta d'usuaris (evolució ambigua). En aquest sentit resulta necessari incorporar noves mesures

que avaluïn qualitat del procés evolutiu de l'aiGA més enllà de la consistència. Concretament, en aquesta tesi s'analitzarà, apart de l'esmentada ambigüitat, l'evolució de l'usuari cap a una o diferents solucions i el nivell de consens entre els diferents usuaris (correlació entre les solucions obtingudes).

1.3.3 Aplicabilitat

Fins a Alías (2006) només s'ha demostrat la viabilitat de l'aiGA en un entorn controlat (8 min. de veu enregistrada) i amb subcostos homogenis: tots ells (*pitch*, energia, durada i coeficients espectrals) s'obtenen de la modelització prosòdica i acústica del senyal. A aquest fet cal afegir que els subcostos no contenen errors provinents de l'etiquetatge del corpus: la parametrització del corpus està supervisada manualment. Aquest escenari s'allunya considerablement d'un escenari real de selecció d'unitats en un CTP-SU de propòsit general. En aquest sentit resulta necessari estudiar l'aplicabilitat de l'aiGA en un entorn real de selecció d'unitats que contempli un corpus més extens ($>$ 1h - corpus inferiors a una hora presenten problemes de selecció ja que contenen unitats amb poca representació (Taylor, 2009)) i incorpori a la funció de cost subcostos heterogenis que l'estat de l'art considera independents (acústics-continus / lingüístics-discrets).

1.3.4 Consens

Quan la mateixa prova d'ajust de pesos utilitzant la metodologia basada en aiGA és realitzada per diferents usuaris, sorgeix el problema de com integrar les diferents solucions que s'han obtingut. Aquest problema s'accentua quan els usuaris discrepen entre ells sobre quina és la millor solució. A Alías *et al.* (2006*a*) aquest problema es solucionava mitjançant una segona validació per part dels usuaris, on les diferents configuracions de pesos obtingudes a partir de la metodologia basada en aiGA seleccionades pels usuaris competien entre elles. En aquest sentit, en la prova de viabilitat resulta necessari un mètode que sigui capaç d'integrar els criteris entre els diferents usuaris, fet identificat pels propis creadors de l'aiGA (Llorà *et al.*, 2008). Malgrat que els diferents usuaris poden discrepar sobre quina és la millor solució al problema, les preferències particulars presentades als usuaris es poden integrar en un únic model global, i així obtenir una solució de consens que satisfaci les preferències dels diferents usuaris. En aquesta tesi aquest problema s'estudia mitjançant models latents, els quals resolen el problema de les contradiccions de manera no heurística.

1.4 Descripció de la tesi

La tesi es divideix en dues parts de recerca principal. A la primera part s'explica l'estat de l'art i la viabilitat d'aplicar ajust de pesos basada en aiGA en CTP-SU, alhora que es fa èmfasi en la col·laboració que es va mantenir durant el seu desenvolupament. A la segona part, es descriuen les noves contribucions realitzades a la metodologia aiGA per tal de millorar-ne la seva eficiència i aplicabilitat. La tesi es complementa amb diversos estudis que analitzen la competitivitat de l'aiGA respecte altres tècniques d'ajust de pesos representatives de l'estat de l'art en un entorn real de selecció d'unitats.

En el capítol 2 es realitza un repàs de la literatura dels CTP-SU i els problemes associats al disseny de la funció de cost juntament amb la investigació realitzada per superar-los. Fruit d'aquesta anàlisi es defineixen les necessitats que ha de complir la funció de cost per tal d'obtenir una veu sintètica d'alta qualitat. En el capítol 3 s'introdueix la capacitat de la computació evolutiva interactiva (IEC) per tal de fusionar esforços humans i computacionals en l'optimització de problemes complexos quan la solució no és coneguda ni lineal. Posteriorment s'expliquen les primeres aproximacions de la IEC per resoldre el problema de l'ajust de pesos en CTP-SU. En el capítol 4 es presenta una nova aproximació en el procés d'ajust perceptiu de pesos basada en la identificació prèvia de clústers d'unitats que adopten pesos semblants en l'ajust automàtic i posterior ajust perceptiu de cada clúster. Aquesta aproximació es porta a terme sense canviar el corpus de treball (*url_fer_ct*) ni els subcostos acústics emprats (Alías, 2006). D'aquesta manera, sense canviar l'escenari d'ajust, es pot estudiar de manera directa la millora aconseguida respecte els treballs previs a aquesta tesi doctoral. Finalment, en el capítol 5 s'aplica la metodologia proposada en un entorn real de selecció d'unitats amb un corpus d'extensió mitjana (*uvig_dav_es* - 1.9h) i subcostos heterogenis (acústics i lingüístics). En aquest capítol, es milloren les contribucions aportades en el capítol 4, i se'n proposen algunes de noves com són canviar el nivell d'ajust automàtic, una nova normalització dels subcostos o millorar l'algorisme de *clustering*. Per finalitzar l'estudi de la tesi doctoral, s'analitza el grau de competitivitat de la metodologia proposada per l'ajust de pesos davant d'altres metodologies d'ajust representatives de l'estat de l'art obtenint resultats prometedors. Al final les conclusions de la tesi es presenten en el capítol 6, així com les línies de futur.

Part I

Fonaments teòrics

CAPÍTOL 2

Conversió de text a parla basada en selecció d'unitats

Els sistemes de conversió text a parla (CTP) basats en selecció d'unitats (CTP-SU) han estat una de les línies d'investigació principal de treball en CTP durant l'última dècada (Cox *et al.*, 2002). Paral·lelament, l'aparició dels CTP basats en models ocults de Markov (HMM) (Tokuda *et al.*, 2003) ha comportat una millora en la flexibilitat d'ús a nivell de requeriments i escenaris, no obstant la naturalitat de la parla sintètica d'aquests sistemes encara no ha assolit la dels CTP-SU quan es disposa d'un corpus extens ($\geq$ 5h de veu). En canvi, sistemes híbrids (CTP-SU+HMM) han aconseguit superar la qualitat de les dues tècniques per separat (Lu *et al.*, 2009). Els sistemes CTP-SU són l'evolució natural dels primers sistemes de síntesi basats en corpus: CTP per difonemes (Moulines i Charpentier, 1990).

Aquest capítol explica amb detall el funcionament dels CTP-SU (Hunt i Black, 1996; Black i Taylor, 1997*a*). Aquest tipus de sistemes es basen en un esquema d'optimització per a seleccionar la millor seqüència d'unitats prèviament enregistrades en un corpus, permetent assolir una millor qualitat global del senyal sintètic. L'esmentat esquema d'optimització (programació dinàmica (Viterbi, 1967)) s'utilitza en moltes aplicacions del processament de la parla, totes elles en l'àmbit del reconeixement i la síntesi (Ostendorf i Bulyko, 2002).

2.1 Introducció

La taxonomia emprada en aquesta dissertació doctoral per explicar els diferents sistemes de síntesi tant des d'una aproximació tècnica com d'una aproximació històrica és la proposada per Taylor (2009), que divideix les tècniques de síntesi en: *i*) tècniques de primera generació (model de tracte vocal), *ii*) tècniques de segona generació (concatenació i modificació mitjançant processament del senyal) i *iii*) tècniques de tercera generació (basades en l'aprenentatge estadístic de veu enregistrada).

2.1.1 Tècniques de síntesi de primera generació

Les tècniques de síntesi de primera generació són les que van dominar l'investigació en la producció de la parla fins a les acaballes del 80. Malgrat que l'ús d'aquestes tècniques és residual avui en dia, el seu estudi serveix per dotar d'una perspectiva històrica la investigació d'aquesta tesi i entendre el perquè els sistemes d'avui dia s'estructuren d'una certa manera. De fet, no es pot entendre el perquè de les tècniques de segona generació, basades en concatenació, sense l'experiència i el coneixement adquirit durant la primera aproximació al problema de la síntesi. És de rigor reconèixer que gran part del mèrit d'obtenir alta qualitat sintètica avui en dia, és gràcies a l'evolució exponencial de les capacitats de la computació en termes de recursos i temps d'execució. Així, en entorns de treball on els recursos són molt limitats, aquests sistemes de síntesi segueixen sent competitius.

Les tècniques de síntesi de primera generació es centren en el modelat del tracte vocal, ja que intenten modelar un sistema que imita el funcionament físic del nostre aparell articulatori, i conseqüentment, generar sons intel·ligibles. A continuació es descriuen molt breument les diferents aproximacions:

1. *Síntesi per formants:* Aquesta aproximació s'inspira en reproduir artificialment l'excitació del pas de l'aire per la glotis així com les ressonàncies per les diferents cavitats de l'aparell fonador humà: nasal, faringe. Això es s'aconsegueix, des d'un punt de vista de processament del senyal, amb la generació d'una font: pels sons sonors mitjançant un tren de deltes (*impulse train*) i una font sorollosa pels sons sords. Aquesta font (*source*) és tractada per un conjunt de filtres ajustats (manualment en la seva primera versió) per produir estimacions acurades de la resposta freqüencial del tracte vocal (formants).

2. *Síntesi per coeficients de predicció lineal (LPC)*: Tècnica derivada de l'anterior que modela la posició dels formants a través d'algorismes de codificació per predicció lineal

aplicats a parla real. La principal diferència és que en comptes de treballar amb un sistema de filtres en paral·lel s'aplica un sol filtre tot pols.

3. *Síntesi articulatòria*: Aquestes tècniques generen la parla modelant directament el comportament de l'aparell fonològic humà. Aquesta síntesi resulta molt dificultosa i gairebé intractable computacionalment parlant, doncs l'aparell fonològic humà resulta molt complex. Anàlogament, aquests sistemes resulten de gran interès a nivell pedagògic.

2.1.2 Tècniques de síntesi de segona generació

La incapacitat de millorar la qualitat robòtica o poc natural d'un senyal inicial produït per un tren de deltes o una font sorollosa va fer palesa la necessitat d'una nova aproximació al problema de la síntesi. Fou necessari abandonar el model simple de deltes/soroll. La segona generació de tècniques de síntesi es caracteritzen per aprofitar-se directament de senyal enregistrat per determinar la veu de sortida, tal com intentava fer la síntesi per LPC. A diferència de l'anterior, l'origen de la forma d'ona es genera mitjançant un senyal de veu autèntic prèviament enregistrat. Malgrat tenir veu enregistrada resulta necessari considerar un model dedicat (explícit) que defineixi quina és l'especificació que ha d'assolir el mòdul de síntesi: les tècniques de primera i de segona generació parteixen de la especificació d'una seqüència de fonemes (al·lòfons) predita per un sistema de conversió text a fonema (*Text-to-Phone* - TTP). La diferència és que un sistema de primera generació generaria el senyal mitjançant un tren de deltes i el posterior mòdul de tracte vocal, mentre que un sistema de segona generació recuperaria el senyal d'una base de dades prèviament enregistrada i li aplicaria algorismes de processament del senyal. L'esmentat senyal seria el corresponent a la seqüència de fonemes especificada.

La gran majoria de sistemes de segona generació treballen amb el difonema com a unitat mínima de representació (Moulines i Charpentier, 1990). El difonema parteix del concepte d'al·lòfon, que s'explica amb més detall a l'apartat 2.2.1 i està conformat per la transició des del centre d'un al·lòfon fins al centre des següent. Així, la paraula "hola" en català està conformada pels següents difonemes: /SIL O/ /Ol/ /l@/ /@ SIL/[1]. Treballar amb aquest tipus d'unitat augmenta la probabilitat d'obtenir concatenacions suaus en concatenar dues unitats diferents degut a que la seva forma d'ona presenta continuïtat (senyal periòdica) en la seva part central. En canvi, concatenar dues unitats diferents per la

[1] En aquest exemple s'ha fet servir la notació SAMPA (Wells *et al.*, 1992) on SIL marca un silenci i @ la vocal neutra.

part inestable del fonema donaria lloc a unions brusques amb salts de fase i de periodicitat deguts, entre d'altres, als efectes coarticulatoris de la parla. No obstant això, és de rigor indicar que la síntesi per difonemes no està exempte d'errors si les marques de periodicitat estan mal alineades.

Una tasca clau en la precisió d'aquests sistemes és l'especificació (intel·ligent) del conjunt de difonemes que conformen les unitats bàsiques a enregistrar (per un idioma) més enllà de la seva combinatòria elemental. A tal efecte s'usen regles fonotàctiques (Sagisaka, 1988), definides a partir de l'estudi de la possibilitat que dos sons puguin aparèixer junts en un determinat idioma o no. Per exemple, com a resultat de l'estudi de regles fonotàctiques en català es conclou que moltes paraules que en llengua estrangera comencen per l'aniuament de consonants necessiten ser adaptades amb l'ajuda d'una vocal a l'inici del mot per fer-lo pronunciable: *slalom* → eslàlom o be *scuola* → escola entre d'altres. Això es veu de manera molt més destacada en les paraules que provenen de llengües indoeuropees, p.ex. *Srbija* → Sèrbia (Prieto, 2004). Malgrat això, el nombre de difonemes descartats degut a la gramàtica fonotàctica és extraordinàriament petit comparat amb el nombre total de sons possibles. Un cop es tanca una llista de combinacions possibles, normalment s'anomena a aquesta llista inventari de difonemes, o en les tècniques de tercera generació, inventari de cobertura (van Santen i Buchsbaum, 1997).

Les diferències entre si dels diversos sistemes de segona generació venen donades per com els algorismes tracten el senyal recuperat de la base de dades. Les tècniques més utilitzades es detallen a l'apartat 2.2.4. Resumidament, aquestes tècniques poden anar des de la simple concatenació amb una simple sincronització de fase (WAV -Beutnagel *et al.* (1998)) fins a tècniques més complexes basats en models sinusoïdals que separen el model harmònic de la parla del model sorollós (Stylianou, 2001).

2.1.3 Tècniques de síntesi de tercera generació

Les tècniques de segona generació, malgrat recuperar la font del senyal de veu d'una base de dades preenregistrada, depenen en excés del disseny de l'inventari de difonemes per mantenir la qualitat en sintetitzar paraules, expressions o frases que no s'han tingut en compte en aquest disseny original. En tots els casos, només es disposa d'una realització d'aquella unitat per a cobrir tots els contextos possibles. En línies generals, les tècniques de síntesi de segona generació només etiqueten, classifiquen, recuperen i processen un senyal de veu ja existent sense aprendre ni establir cap model més enllà dels contextos i vocabularis (dominis) prèviament enregistrats.

És per aquest motiu que en els últims anys han aparegut les tècniques de síntesi de tercera generació. Aquestes tècniques pretenen construir un model de producció de la parla que resolgui satisfactòriament la síntesi de la parla tant en dominis dissenyats prèviament (limitats) com en dominis desconeguts (oberts o de propòsit general). Aquesta obertura és gràcies a modelats estadístics que no només recuperen les dades enregistrades sinó que són capaços d'inferir certa regressió o predicció de casos no considerats en el disseny original. Per assolir les necessitats més elementals i alhora aprendre del procés natural de producció de la parla s'empren una gran varietat de tècniques d'optimització i aprenentatge artificial (en el seu sentit més pur). Són aquestes tècniques les que permeten fer un salt endavant i passar de les tècniques de segona generació a les de tercera generació. En altres paraules, mentre les tècniques de segona generació memoritzen, les tècniques de tercera generació aprenen (Taylor, 2009).

Les tècniques de síntesi de tercera generació es poden dividir a la vegada en tècniques basades en models ocults de Markov (*Hidden Markov Models* - HMMs -Tokuda *et al.* (2003)) i tècniques basades en selecció d'unitats (CTP-SU - Hunt i Black (1996)). La principal diferència entre elles és que mentre els HMM realitzen un aprenentatge estadístic que dota al sistema de solidesa i flexibilitat, els sistemes CTP-SU optimitzen un procés per a poder recuperar la millor unitat enregistrada (funcionen com els sistemes de segona generació però d'una manera intel·ligent) i per tant, aporten més qualitat en detriment de la flexibilitat i la solidesa del model. Els motius de l'idoneïtat de la solidesa o la flexibilitat en funció de la qualitat de la veu sintètica s'expliquen a continuació.

Models ocults de Markov

Els HMMs empren les unitats enregistrades com a un conjunt d'observacions per l'aprenentatge. La parametrització comuna es serveix de la informació següent: coeficients cepstrals a l'escala Mel (*Mel-Frequency Cepstrum Coefficients* - MFCCs), la freqüència fonamental juntament amb les seves variacions (deltes) i, en alguns casos, informació addicional de la font. Tota aquesta informació es parametritza en funció del context (Tokuda *et al.*, 2003). L'aprenentatge primer aplica algorismes d'esperança maximització (*Expectation Maximization* - EM) per trobar seqüències d'informació, per trams, de les unitats i contextos específics. Fruit d'aquest entrenament s'obtenen vectors composats per coeficients d'acceleració i les corresponents deltes. Aquests vectors són la representació parametritzada del senyal. L'entrenament es realitza per cada combinació possible de paràmetres d'entrada (unitats i contextos). Per tal d'agrupar casos amb poca representació s'empren arbres de decisió contextuals. Aquest tipus d'aprenentatge, aconsegueix models compactes, de fàcil

modificació a efectes de transformació de veu i altres propòsits (Tokuda *et al.*, 2003).

L'algorisme EM (Dempster *et al.*, 1977) aplica el principi de màxima versemblança (*Maximum Likelihood* - ML) que controla el biaix de l'entrenament a unes dades concretes (sobreaprenentatge, subaprenentatge) així com permet obtenir un modelat robust respecte dades incompletes o incorrectes. Així es dota al sistema d'una consistència sòlida més enllà del seu conjunt d'entrenament i a la vegada es proporciona una alta flexibilitat per a poder sintetitzar veu fora del domini de contextos i vocabulari prèviament enregistrats.

Tot i així, tal com s'ha esmentat anteriorment, malgrat l'augment de la flexibilitat fora del domini, aquestes tècniques, no han aconseguit a hores d'ara superar la qualitat dels sintetitzadors basats en selecció d'unitats quan aquests últims disposen d'un corpus extens i ben dissenyat ($\geq$ 5h), si bé que les tècniques híbrides de selecció d'unitats guiades mitjançant parametritzacions amb models ocults de Markov són les que obtenen millor resultat en l'actualitat (Oura *et al.*, 2009; Lu *et al.*, 2009).

Selecció d'unitats

L'altra tècnica de tercera generació és la síntesi concatenativa per selecció d'unitats (CTPSU), que és la tècnica de síntesi d'alta qualitat dominant avui en dia. Aquesta tècnica és la derivació natural de la síntesi per difonemes (veure apartat 2.1.2)

La selecció d'unitats es basa fonamentalment en enregistrar i recuperar més d'una realització per cada unitat (normalment difonema, tot i que la unitat pot ser de mida variable). Això permet disposar d'un origen del senyal que no ve determinat, de manera fixa, per coeficients tancats. En disposar de senyal natural es poden tenir les diferents versions de la mateixa en funció del context de manera natural.

L'èxit d'aquesta tècnica es basa en la no restricció, a diferència dels sistemes de segona generació, les dues assumpcions implícites que es presenten a continuació: (Taylor, 2009):

1. Dins d'un tipus de difonema, es poden assolir totes les variacions possibles variant el *pitch* (apartat 2.2.2) i la durada amb tècniques de processament del senyal.
2. Els algorismes de processament del senyal són capaços de realitzar totes les modificacions de *pitch* i durada sense cap pèrdua de naturalitat.

El senyal de veu té característiques acústiques complexes, no totes conegudes, més enllà del seu ritme (velocitat), energia i freqüència fonamental (paràmetres que conformen la prosòdia clàssica com es detalla en l'apartat 2.2.2). Alguns estudis demostren que un

mateix difonema pot tenir idèntics valors de ritme i freqüència fonamental, però sonar diferent segons si es troba en una síl·laba accentuada o no, si és final de frase o si és principi de paraula (Dutoit, 1997; Cabral i Oliveira, 2006). És per aquest motiu que implícitament s'ha de considerar tota aquella informació que desconeixem enregistrant més d'una versió per unitat, i així disposar d'informació del senyal que no s'ha tingut en compte prèviament. Exemples d'aquesta informació, coneguda o no, són la qualitat de veu en entorns expressius (Cabral i Oliveira, 2006) o la coarticulació (Dutoit, 1997) entre d'altres.

Els sistemes CTP-SU es defineixen per un conjunt de tècniques que s'apliquen en diferents parts del procés de síntesi de la parla per tal de superar aquesta complexitat. Aquestes tècniques s'apliquen en diferents etapes del procés que van des de l'anàlisi lingüístic i parametrització prosòdica, fins en l'àmbit de concatenació i modificació del senyal, passant lògicament per una fase específica de selecció de les millors unitats.

En teoria el problema es podria reduir a simplement augmentar el vector d'especificacions inicials (entrada dels sistemes de segona generació). És a dir, en comptes d'enregistrar una realització per cada difonema obtingut per regles fonotàctiques, enregistrar un difonema per cada tipus d'especificat contextual, valors de durada i freqüència fonamental (Strom *et al.*, 2006; Cui *et al.*, 2007). Malgrat que es pot pensar d'enregistrar un corpus de veu segons aquestes característiques hi ha dues principals dificultats que fan més raonable una altra aproximació al problema.

La primera dificultat d'aquest enfoc seria la tasca de disseny d'aquest corpus. Aquest fet suposaria una tasca dificultosa obtenir un inventari de l'ordre, per exemple en català, de 60.000 difonemes dissenyats un a un per ser enregistrats segons el seu context. La segona dificultat seria enregistrar aquests sons aïlladament. Encara que disposéssim del millor locutor professional disponible, enregistrar únicament dos sons aïllats i alhora lleugerament diferents en termes de coarticulació suposa una tasca que va més enllà de les tècniques de locució (Lenzo i Black, 2000). És en aquest paradigma que apareix la idea d'emprar frases portadores: s'enregistrarien un conjunt de frases que contindrien, de manera explícita, el difonema dissenyat. No obstant, aquesta aproximació comporta una feina extra de postproducció a l'haver de separar els difonemes dissenyats dels difonemes comuns dins la frase portadora. Aquesta última tasca extra a priori sembla inútil, ja que si s'aprofités tota la frase, etiquetada correctament, no empitjoraria en cap cas la qualitat final: simplement es disposaria de més redundància d'uns difonemes respecte dels altres.

És en aquest context que s'obté la metodologia disseny i enregistrament àmpliament acceptada: s'agafen una sèrie de textos dins d'un o diversos dominis (en el cas que es vulgui fer un sintetitzador de propòsit general) i es seleccionen aquelles frases que aporten

més variabilitat fonètica i contextual (Strom *et al.*, 2006; Cui *et al.*, 2007). Posteriorment es completa l'inventari, si és necessari, amb aquells sons que rarament apareixen o són escassos en un idioma. Aquestes paraules són les que es coneixen com a paraules de cobertura (van Santen i Buchsbaum, 1997). Llavors, s'utilitza tota aquesta veu enregistrada com a corpus per a poder realitzar la selecció d'unitats. D'aquesta manera, la síntesi concatenativa esdevé una tècnica basada en l'aprenentatge de locucions reals en comptes d'una tècnica de disseny lingüístic basat en regles: esdevenint així una tècnica de tercera generació.

L'estudi en la síntesi a través de selecció d'unitats considera que la qualitat òptima només s'assoleix si es selecciona allò millor minimitzant les modificacions "*choose the best and modify the least*" (Balestri *et al.*, 1999): concloent que quan més processament matemàtic es fa al senyal de veu, pitjor qualitat i naturalitat s'obté.

En el proper apartat s'explica l'estructura d'aquests tipus de sistemes de síntesi amb més detall.

2.2 Estructura dels sistemes CTP-SU

Els CTP-SU típicament s'estructuren en quatre fases (veure figura 2.1). La primera fase és l'anàlisi mitjançant processament del llenguatge natural, la segona la generació de la prosòdia de la parla a sintetitzar, la tercera fase és la selecció d'unitats i la quarta fase és la generació de la forma d'ona.

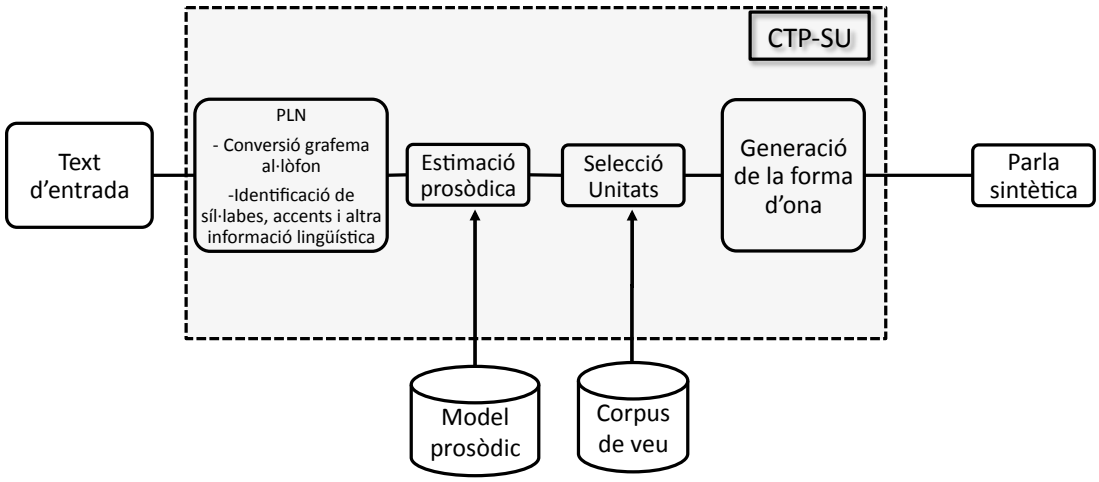


Figura 2.1: Etapes d'un sistema de conversió de text a parla basat en selecció d'unitats.

2.2.1 Anàlisi del text

La primera fase és l'anàlisi del text mitjançant el processament del llenguatge natural. Aquest anàlisi, principalment, consta de tres fases: normalització, segmentació (que alhora es divideix entre anàlisi lèxic -*tokenizing*- i separació en grups entonatius -*phrasify*-), conversió de grafema a al·lòfon (transcripció fonètica) i, per últim, generació d'informació morfològica (*Part-of-Speech* (Clark *et al.*, 2007)), pausal i accentual.

Normalització

La normalització del text és aquell procés pel qual un text passa d'una representació general escrita a una representació canònica (sense ambigüitats). En contextos de síntesi de propòsit general l'origen del text pot tenir orígens diferents: una web (p.ex. text basat en etiquetes SGML), un document de treball (p.ex. processador de textos, fulla de càlcul), un text etiquetat segons un format específic (p.ex.LaTeX, *Postscript* a dos columnes), un text obtingut de processos de reconeixement de la parla (p.ex. ViaVoice, VoiceOver, Loquendo), sistemes de diàleg o de traducció automàtica entre d'altres. És per aquest motiu que el text d'entrada pot ser heterogeni i necessita ser normalitzat a un format concret, sigui quin sigui el seu origen, a la sortida hi hagi el mateix format. Normalment, aquesta normalització implica tasques de: *i*) categorització dels signes de puntuació, *ii*) expansió d'abreviatures i acrònims (Monzo *et al.*, 2006), *iii*) eliminació de parts supèrflues (cites, notes a peu de plana, etiquetes de format del text, etc.), *iv*) conversió de dígits a text (p.ex. dates, nombres, moneda) i *v*) detecció d'estrangerismes i excepcions en la parla que necessiten o bé ser reparades, o bé etiquetades d'una manera especial. La informació obtinguda després d'aquest procés s'anomena representació canònica del text d'entrada.

Segmentació

La segmentació s'encarrega d'obtenir parts mínimament significants dividint el text canònic d'entrada i estableix relacions entre elles. En totes les seves etapes s'ajuda dels signes de puntuació. Primerament realitza un anàlisi lèxic (*tokenizer*) i determina quin conjunt de símbols (grafemes, guions, apòstrofs, etc.) conformen una unitat significativa coneguda com a unitat lèxica o *token* (normalment una paraula). Un cop s'han determinat les unitats lèxiques, es delimita el grup d'entonació o *utterance*. Aquest procés, que en anglès rep el nom de *phrasing* determina quines relacions tenen les unitats lèxiques entre sí. Una frase delimitada ortogràficament generalment conforma un o varis grups d'entonació. Si

s'aplica un model de pausat, aquest delimita on s'han d'incorporar els silencis en el text.

Conversió de lletra a so

El tercer procés es basa en convertir cada grafema (o conjunt de grafemes) a un o un conjunt d'al·lòfons (a poder ser s'evita que siguin fonemes). Cal no confondre el concepte fonema (*phoneme*) del concepte al·lòfon (*phone*). Tot i que l'anàlisi amb detall de la representació simbòlica del llenguatge queda fora de l'àmbit d'aquesta tesi, el concepte d'al·lòfon és important per a la posterior selecció d'unitats. Un fonema representa un so concret dins d'un idioma dotat de seu significat mínim. Però tanmateix, un al·lòfon és la manera exacta de reproduir un so de la mateixa manera per a tothom. Per exemple, en català pel mateix fonema nasal /n/ es pot velaritzar coarticulant-se fortament amb la /k/ (encabir → IPA: /əŋkəβˈi/ SAMPA:/@Nk@B'i/) o bé es pot coarticular amb un so fricatiu esdevenint així labiodental (enfadar → IPA:/əɱfəðˈa/ SAMPA:/@Ff@D'a/).

Generació de la informació contextual, entonativa i pausal

L'última etapa de l'anàlisi lingüístic es divideix en dues anàlisis en paral·lel relacionades entre sí. Cal diferenciar una etapa d'anàlisi morfològic basada en associar cada unitat lèxica a una categoria gramatical (*Part-of-Speech*) estructurant les unitats lèxiques en síl·labes (generalment morfemes) i una segona etapa de generació d'informació pausal i d'entonació. En aquesta segona etapa s'especifica, per una banda, quines síl·labes estan accentuades i segons l'idioma quin tipus d'accent tenen (informació accentual), i per altra banda, també quines pauses s'han de fer juntament amb el seu tipus, que determina la seva durada.

Mentre que l'anàlisi morfològic segueix un patró bastant semblant per tots els idiomes (separació de la paraula en lexema i morfema, consulta d'un diccionari gramatical i obtenció d'informació de categoria gramatical, gènere, nombre, temps...) la informació entonativa i pausal segueix patrons de comportament completament diferents per cada idioma o fins i tot variant dialectal. Per exemple, en anglès americà aquesta informació generalment es representa mitjançant les regles ToBI (*Tonal and Breaks Indices*), en canvi l'anglès britànic separa aquesta informació en etapes (*stages*) i assigna una forma d'ona per cada etapa. En castellà i català, generalment, la informació entonativa s'estructura en posició dins grups accentuals, posició dins grups d'entonació i tipus d'enunciació (interrogativa, exclamativa o enunciativa), sense tenir, a priori, una estructura específica definida. La entonació pot denotar la intencionalitat de la frase, o quina part de la frase té més èmfasi: no és el ma-

teix dir *ahir **en Felip**, va anar al teatre musical* que dir *ahir, en Felip va anar al **teatre musical***) (Bulyko, 2002). En altres idiomes, l'entonació fins i tot pot denotar la semàntica de la paraula (p.ex. la paraula en xinès mandarí *ma* pot significar mare, renyar, cavall o cànem en funció de la seva entonació).

La parametrització ToBI és util, entre moltes altres coses, pel seu modelat de les pauses. ToBI defineix quatre tipus de silencis (Beckman i Hirschberg, 1994): silencis de principi i final de locució, silencis forts entre frases (delimiten una grup d'entonació-*utterance* d'altres grups d'entonació), silencis dèbils dins la frase (no trenquen el grup d'entonació, per exemple, coma i punt i coma) o silencis espontanis causats o bé per difluències (Adell *et al.*, 2006) que introdueixen certa pseudonaturalitat dins la parla o bé per marcar un èmfasi.

2.2.2 Estimació de prosòdia

Un cop s'ha realitzat l'anàlisi lingüístic, ja es disposa del text d'entrada segons una representació que és independent de l'idioma (canònica). D'aquesta informació se'n diu notació simbòlica. La notació simbòlica es pot convertir a una especificació contínua de paràmetres acústics que parametritzen un senyal del veu. D'aquesta especificació contínua (Bulyko, 2002) se'n diu prosòdia. La prosòdia es defineix com el conjunt de l'al·lòfon amb la seva informació segmental bàsica (contorn de la freqüencial fonamental o *pitch*, energia i durada). Cal afegir que autors previs a la selecció d'unitats (Campbell, 1990) consideren la síl·laba com a unitat mínima de prosòdia. Altres característiques de la prosòdia poden ser paràmetres de qualitat de veu (*Voice Quality* - VoQ), paràmetres de la font d'algunes senyals o d'altres. La informació prosòdica és molt important per la intel·ligibilitat, l'expressivitat i la naturalitat de la parla.

El modelat de la prosòdia, sobretot el modelat de la informació entonativa, no segueix un sol proces predeterminat coexistint així diverses aproximacions al problema que poden, dependre de l'idioma.

A Iriondo (2008) es realitza un repàs profund de la generació de prosòdia en català i en castellà, que es passa a descriure a continuació: en català i castellà el model d'entonació bàsic ve determinat pel contorn de la F_0 a través d'un grup d'entonació (Garrido, 2001). A Escudero i Cardeñoso (2003) es fa una revisió dels diferents models que caracteritzen la entonació, també dita melodia. En aquell treball es detalla que per modelar el contorn melòdic es poden considerar diferents segments: *i*) el concepte de micromelodia o models de Fujisaki (Fujisaki, 1992), que associa el contorn de la corba de F_0 dins els límits de la síl·laba, *ii*) considerar l'entonació dins del grup accentual que normalment es relaciona

amb el ritme de la parla, i *iii*) el grup d'entonació i d'altres unitats superiors, que només intervenen a l'hora d'elaborar un discurs sigui per fer èmfasis.

El CTP en el qual s'emmarca aquesta tesi segueix la proposta de modelat prosòdic de Iriondo (2008) en la que es proposa treballar amb un model de F_0 per al·lòfon els quals segueixen els punts d'una corba (aproximació polinòmica) que té els seus punts d'inflexió en cada vocal tònica. Per tant, el grup accentual és la unitat bàsica pel modelat de la melodia considerant aquest com a una paraula accentuada precedida, si és el cas, per una o més paraules àtones. Per més informació es pot consultar la tesi d'Iriondo (2008).

D'altra banda, el modelat de les durades de fonemes és bastant homogeni en tots els idiomes i, en general, es calcula com la variabilitat de la pròpia durada de l'al·lòfon en funció de la mitjana i la desviació típica de cadascun d'ells dins del corpus. Aquesta normalització és coneguda com a normalització *z-score* (Klatt, 1979; Navas *et al.*, 2002*a*; Schweitzer i Möbius, 2004).

Històricament, en una primera aproximació del modelat prosòdic, la prosòdia es modelava amb coneixement expert, sense cap tipus d'aprenentatge estadístic. Aquests models s'empraven en els sistemes de síntesi de primera i segona generació per tal de proporcionar prosòdia al procés de síntesi. Alguns exemples d'aquests sistemes els podem trobar en el model acústic de predicció de durades de Klatt (Klatt, 1979) o el model de predicció entonativa de Fujisaki (Fujisaki, 1992), tot i que aquest últim s'havia obtingut d'un anàlisi estadístic supervisat d'un corpus oral de lectura i diàleg. Fins i tot en el treball de Sagisaka (Sagisaka, 1988), que fou preludi d'un sistema CTP-SU tal com el coneixem avui, es pot observar un model prosòdic basat en regles. El modelat ToBI (Beckman i Hirschberg, 1994) en la seva primera versió també estava basat en un sistema de regles.

A mitjans dels anys 90, amb l'aparició dels primers sistemes de selecció d'unitats, es van emprar arbres de decisió i regressió (*Classification and Regression Trees* - CART) (Mobius i Von-Santen, 1996; Bailly *et al.*, 1992) per a determinar la durada i freqüència de cada al·lòfon. Aquests arbres de decisió recuperaven la informació lingüística i contextual detallada en l'apartat anterior, i a base d'aprenentatge estadístic definien una parametrització prosòdica a ser assolida pel mòduls posteriors en el sistema.

Altres estudis de modelat prosòdic inclouen el raonament basat en casos (CBR) (Iriondo *et al.*, 2007), xarxes neuronals (Campbell, 1990), xarxes Bayesianes (Goubanova i Taylor, 2000), entre d'altres.

Dins del treballs de recerca actuals en modelat prosòdic cal fer una menció especial al treball de Campillo i Rodríguez-Banga (2006); Clark i King (2006) els quals no dissocien el

fet de predir el contorn d'entonació a la l'existència del mateix en el corpus. Per tant, per a cada síntesi, el mòdul de prosòdia retorna una llista dels n millors contorns d'entonació que serviran per guiar la selecció d'unitats. El patró d'entonació que es farà servir al final ve determinat per les unitats que trobi el mòdul selector d'unitats, i si aquestes satisfan un o altre contorn d'entonació.

També cal considerar, com es detalla en el proper apartat, que hi ha sistemes que no realitzen una selecció d'unitats subordinada a la predicció prosòdica de valors acústics de manera contínua, sinó que guien la selecció d'unitats directament amb la notació simbòlica obtinguda a partir de l'anàlisi lingüístic (Clark *et al.*, 2007; Bulyko, 2002). Aquests sistemes consideren que la informació acústica de baix nivell que dóna continuïtat a la prosòdia es troba implícita en la informació paramètrica obtinguda de l'anàlisi lingüístic. Per tant, ells consideren que si es pot realitzar la selecció d'unitats en funció d'aquests paràmetres s'evita una etapa que pot afegir soroll al procés.

Per últim, anomenar que algunes aproximacions híbrides (HMM+CTP-SU) combinen la predicció prosòdica amb aprenentatatge HMM i selecció d'unitats (Toda, 2003). Aquests sistemes fan servir arbres de decisió per predir les durades de les unitats i un cop definida la durada, es separa la informació de l'al·lòfon en tres etapes (transició inicial, part estable i transició final) i llavors es realitza una predicció de F_0 i valors MFCC per a cada etapa. Llavors, amb aquesta informació es guia també la selecció d'unitats, realitzant per últim una síntesi concatenativa ajustant la prosòdia original amb la tècnica coneguda com *STRAIGHT* (Toda, 2003).

2.2.3 Selecció d'unitats

El procés que ha comportat el salt qualitatiu de la segona a la tercera generació en la branca de la síntesi concatenativa ha estat el desenvolupament del mòdul de selecció d'unitats, creant així un nou paradigma en els conversors de text a parla. A dia d'avui, aquest mòdul juntament amb un bon corpus és el que permet obtenir una millor naturalitat, riquesa i expressivitat en la producció de la parla automàtica.

La idea bàsica és que, mitjançant la selecció d'unitats, es seleccioni un conjunt òptim d'unitats d'un corpus de veu prèviament enregistrat. Aquesta optimització s'assoleix minimitzant una funció que avalua la degradació de la naturalitat. Aquesta degradació ve causada per: *i*) la diferència prosòdica, *ii*) la diferència espectral i *iii*) desalineament en els entorns fonètics (Sagisaka, 1988).

Punt de partida

A finals dels anys 80 i a principis dels anys 90, sobretot en l'àmbit de la recerca dels CTP en japonès, van sorgir diverses propostes per a realitzar una síntesi amb més d'una versió enregistrada per unitat (Sagisaka, 1988; Nakajima i Hamada, 1988; Sagisaka *et al.*, 1992; Iwahashi *et al.*, 1993). Tanmateix, no va ser fins el 1996 quan Andrew Hunt i Alan W. Black van proposar una arquitectura general al problema de la selecció d'unitats després de molts anys de recerca en aquest sentit (Hunt i Black, 1996). Van assentar les bases d'un procés amb un article que ha esdevingut el punt de partida per qualsevol treball de recerca en l'àmbit la selecció d'unitats.

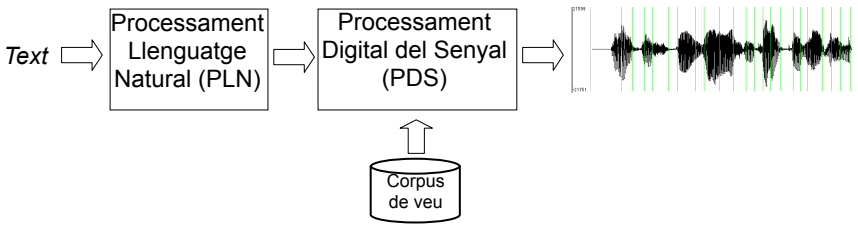


Figura 2.2: Diagrama de blocs de la síntesi per difonemes.

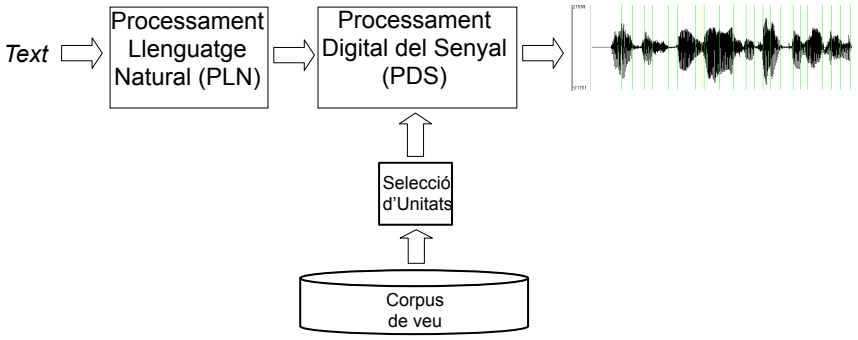


Figura 2.3: Diagrama de blocs de la síntesi per selecció d'unitats.

Obtenció de la unitat adequada per a l'especificació desitjada: la funció de cost de la selecció d'unitats

En aquest paradigma clàssic (Hunt i Black, 1996), es defineix la selecció d'unitats com un problema d'optimització (cerca) a través de totes les possibles seqüències d'unitats candi-

dates per trobar la millor seqüència en termes de naturalitat. El concepte de naturalitat és ambigu, tanmateix, en el context de la selecció d'unitats, queda definida mitjançant una funció de cost global composada a la vegada de diferents funcions de subcost. A l'hora de combinar diferents subcostos, Hunt i Black segueixen l'enfoc clàssic per a solucionar una optimització multiobjectiu inspirada en el concepte de distància de Manhattan ponderada (Srinivas i Deb, 1994). El procés consisteix en assignar un pes w_i a cadascun dels objectius normalitzats SC_k per a convertir el problema en un problema d'un sol objectiu. Aquests objectius SC_k son les diferents funcions de subcost que avaluen la degradació de la naturalitat segons diferents aspectes tal i com es detalla en l'apartat 2.3.3. Així la funció de cost global queda definida com una distància de Manhattan ponderada (funció objectiva escalar) tal com es pot veure en l'equació 2.1.

$$C(i,j) = w_1 SC_1(i) + w_2 SC_2(i) + \cdots + w_k SC_k(i,j) \tag{2.1}$$

En la notació de l'equació w representa el pes que pondera el subcost normalitzat SC, on els seus valors són de l'interval [0,1] i $\sum w_i = 1$. $C(i,j)$ representa l'elecció de las unitats candidates i i j

Aquest enfoc és una aproximació a priori ja que requereix que els pesos siguin coneguts. El fet de solucionar un problema amb una funció objectiu per a un vector de pesos donat $w_i = \{w_1, w_2, \ldots w_k\}$ deriva cap a un sol resultat (que serà una seqüència d'unitats). Per tant, com que no són necessàries múltiples possibilitats, no s'observa en cap moment la necessitat d'atacar el problema amb un enfoc multiobjectiu clàssic (per exemple, basat en fronts de Pareto (Horn *et al.*, 1994)). Tot i així, la dificultat principal d'aquest enfoc és trobar el vector de pesos apropiat, tema central d'aquesta tesi que s'estudia a fons en l'apartat 2.4.

Les diferents funcions de subcost modelen, cadascuna d'elles, un cost per a cada paràmetre acústic o simbòlic de les unitats. Aquestes funcions es poden agrupar segons un subcostos objectiu o de *target* (eq. 2.2) i subcostos de concatenació (eq. 2.3). Els subcostos de *target* (C_T) quantifiquen la diferència de les unitats seleccionades respecte una especificació prosòdica o lingüística original i els subcostos de concatenació (C_C) quantifiquen la diferència de la unió dels diferents trams del senyal de veu en el seu punt de concatenació. La notació seguida pels subcostos és SC.

Llavors, formalment la funció de cost de selecció de la unitat i juntament amb la unitat j queda definida per:

$$C_T(i) = \sum_{k=0}^{param\_t} w_i^k \cdot {SC_T}^k(i) \tag{2.2}$$

$$C_C(i,j) = \sum_{k=0}^{param\_c} w_{ij}^k \cdot {SC_C}^k(i,j) \tag{2.3}$$

$$C(i,j) = C_T(i) + C_C(i,j) \tag{2.4}$$

on ${SC_T}^k(i)$ i ${SC_C}^k(i,j)$ representen respectivament, els subcostos de *target* i el de concatenació, i es calculen com:

$${SC_T}^k(i) = D\left[P(u_i)^k, P(t_i)^k\right] \tag{2.5}$$

$${SC_C}^k(i,j) = D\left[P(u_i^R)^k, P(u_j^L)^k\right] \tag{2.6}$$

on u_i és la unitat candidata, t_i és la especificació de la unitat, u_i^R és la parametrització en el punt de concatenació (dret) de la unitat candidata i (esquerra) i u_j^L és la parametrització en el punt de concatenació (esquerre) de la unitat candidata j (dreta). $D[\cdot]$ és la funció de distància (Manhattan, euclídea, cúbica...) i $P(\cdot)_k$ és el valor mesurat del paràmetre k en la unitat.

Existeixen moltes variants que seran analitzades més endavant sobre la naturalesa dels subcostos a emprar (apartat 2.3)

Obtenció de la millor seqüència d'unitats mitjançant programació dinàmica

Un cop definida una funció de cost que quantifica la idoneïtat de les unitats candidates per ser seleccionades resulta necessari un nou pas. Aquest nou pas és l'obtenció de manera intel·ligent de la seqüència d'unitats que minimitza la funció de cost global. Segons la definició clàssica (Hunt i Black, 1996) l'algorisme de Viterbi (Viterbi, 1967) és el mètode que realitza millor aquesta etapa.

L'algorisme de Viterbi segueix les bases del paradigma de programació dinàmica dissenyat per Richard Bellman (Bellman, 1954). Segons aquests principis, un problema gran i complex es pot subdividir en problemes de solucions localment òptimes, per tant la solució global s'obté mitjançant la composició d'aquestes solucions òptimes dels subproblemes derivats del problema original.

Una selecció d'unitats òptima ha d'obtenir el millor camí (seqüència única d'unitats) amb el mínim cost acumulat. Per assegurar el mínim absolut s'haurien de comprovar totes les seqüències d'unitats candidates possibles. Tanmateix, el nombre de seqüències possible pot arribar a ser extraordinàriament gran i és aquí quan el principi de composició

de la programació dinàmica i l'algorisme de Viterbi cobren importància: permeten trobar el millor camí possible segons un temps d'execució lineal en comptes d'un temps exponencial (Viterbi, 1967).

Algorisme 2.1 Etapes de l'algorisme de Viterbi.

Inicialització

for $i = 1$ to N **do**
  $\delta_1 i = C_T(i)$
end for

Recursivitat

for $t = 1$ to $T - 1$ **do**
  for $j = 1$ to N **do**
    $\delta_{t+1}(i) = \left[\min_{1 \leq i \leq N} \delta_i(C_C(i,j)) \right] C_T(j)$
  end for
end for

Acabament

for $t = 1$ to $T - 1$ **do**
  for $j = 1$ to N **do**
    $\varphi_{t+1}(j) = \arg\min_{1 \leq i \leq N} \delta_t(i) C_C(i,j)$
  end for
end for
$s_T^* = \arg\min_{1 \leq i \leq N} (\delta_T(i))$

Reconstrucció de la seqüència d'estats amb un mínim cost

for $t = T - 1$ to t **do**
  $s_t^* = \varphi_{t+1}(s_{t+1}^*)$ { s_t^* representa la seqüencia òptima}
end for

Clàssicament el procés de selecció d'unitats es pot representar com un diagrama de Trellis (Bahl *et al.*, 1974) de N estats (difonemes candidats en el cas de la síntesi) per T observacions (mida de la seqüència de difonemes). Llavors s'entén el diagrama de Trellis com un diagrama de T etapes temporals. A cada etapa temporal, es selecciona una sola unitat entre les seves N unitats candidates, i per cada estat s'estableixen transicions entre l'etapa temporal t i la $t + 1$, per $t < T$. La cerca explora cadascun dels estats, calculant els subcostos a mesura que s'avança en el diagrama (procés *forward*). Un cop explorats tots els estats, es reconstrueix la seqüència des del final cap al principi tinguent en compte el millor camí que arriba a l'estat, per així reconstruir la seqüència òptima (procés *backward*).

Així doncs, l'algorisme permet trobar la seqüència d'unitat que adopta un mínim cost

en un diagrama de Trellis, $S = (s_1, s_2, \ldots, s_T)$ essent s_i la posició d'una unitat en un instant determinat a partir d'una especificació d'unitats prèvia $U = (u_1, u_2, \ldots, u_T)$, és a dir, obté la seqüència òptima que millor satisfà la seqüència especificada.

A l'algorisme 2.1 es detalla el procés complet a través de les funcions δ i φ. La funció $\delta_t(i)$ representa el cost del millor camí fins la unitat candidata i havent explorat les t primeres unitats. $\delta_t(i)$ es calcula per tots els estats i instants de temps. L'esmentat camí s'indexa amb un únic identificador (argument). Es memoritza $\varphi_t(j)$. $\varphi_t(j)$ per recordar el camí traçat en l'etapa de reconstrucció (*backward*), degut que l'objectiu és recuperar la seqüències d'estats amb un mínim cost.

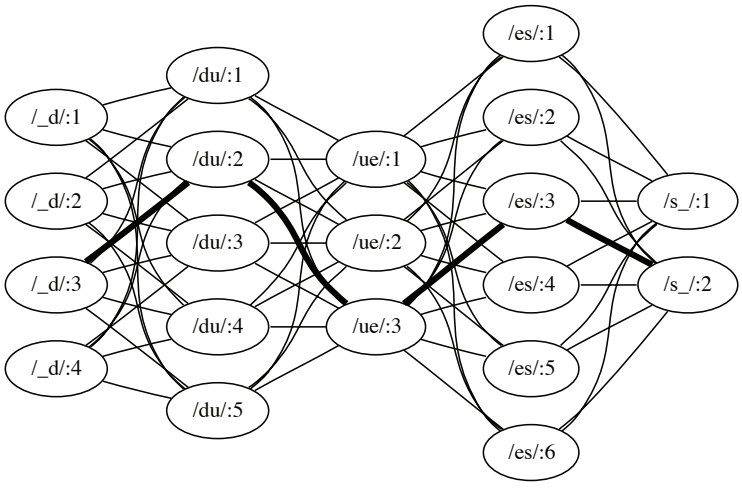


Figura 2.4: Exemple del resultat d'aplicar l'algorisme de Viterbi per a la selecció de les millors unitats que conformen la paraula dues, /SIL-d//d-u//u-e//e-s//s-SIL/ sobre una estructura Trellis de mida $N \times T$, on $N = \{4, 5, 3, 6, 2\}$ i $T = 5$. La línia gruixuda indica la millor seqüència d'unitats.

Alguns dels càlculs de l'algorisme de Viterbi recorden als de l'algorisme *forward* BCJR (Bahl, Cocke, Jelinek i Raviv, 1974) plantejat inicialment per resoldre diagrames de Trellis. La diferència fonamental és la incorporació de la funció argmin (en comptes de sumar valors dels camins) per a calcular la seqüència d'estats amb un mínim cost. Un exemple de l'algorisme de Viterbi sobre el diagrama de Trellis de la paraula "dues" es pot observar a la figura 2.4.

2.2.4 Generació de la forma d'ona

Un cop s'ha seleccionat la millor seqüència d'unitats del corpus de veu, ja es disposa dels segments de veu que conformaran la síntesi final. L'última etapa és l'encarregada de concatenar els diferents segments que donaran lloc al senyal sintètic final. Per tal de realitzar aquest procés existeixen diverses tècniques, la complexitat de les quals va des de simplement alinear el senyal sense realitzar-hi cap tipus de modificació (WAV - Beutnagel *et al.* (1998)) fins a una descomposició del senyal en models harmònics i de soroll, obtenint models específics per les parts sonores i sordes (HNM -Stylianou (2001)). Aquesta descomposició permet la seva transformació en termes de continuitat espectral i poder-lo tornar a composar.

La tria entre una o altre tècnica de processament del senyal i concatenació pot venir donada per la bondat de les unitats obtingudes a través del procés de selecció d'unitats. En el cas utòpic d'una selecció d'unitats perfecta, la simple concatenació del senyal seria idònia sense la necessitat d'aplicar-hi qualsevol tipus de processament: aquest deterioraria la qualitat final dels segments de veu obtinguts. Per alinear en fase diferents segments de veu s'empren tècniques simples de superposició i suma síncrones segons el *pitch* (*overlap and add*).

A continuació s'analitzen les diferents tècniques de concatenació i posterior transformació del senyal classificades segons el seu principi de funcionament. La gran majoria es van desenvolupar durant la síntesi de segona generació basada en difonemes i processament del senyal.

Les principals tècniques per tractar el senyal recuperat en síntesi concatenativa són les següents(Taylor, 2009):

1. **Concatenació simple (Concatenació RAW o WAV)**: Concatena el senyal de les unitats de manera directa amb una mínima alineació de fase (Beutnagel *et al.*, 1998).

2. **Predicció lineal de l'excitació residual**: Primer es realitza una anàlisi per predicció lineal de tot el senyal de veu, incloent també la part residual per la resíntesi en comptes de fer servir només els impulsos (Hunt *et al.*, 1989; Vincent *et al.*, 2005).

3. **Models sinusoïdals**: Empren models harmònics i descomponen les trames en conjunts d'harmònics d'una freqüència fonamental estimada. Els paràmetres del model són les amplituds i les fases dels harmònics. Amb aquests es pot canviar el valor de la freqüència fonamental mantenint la mateixa envolupant espectral (Macon, 1996).

4. **Models harmònics amb soroll**: Són semblants als models sinusoïdals, però a diferència dels anteriors adopten un model de soroll que permet refinar el modelat dels intervals de senyal sonors amb una alta freqüència de comportament sorollós o totes les parts de parla sorda (Stylianou, 2001).

5. ***Time-Domain Pitch Synchronous OverLap and Add* (PSOLA)**: Mètode de concatenació en el domini temporal. Separa el senyal de veu original en trames *pitch*-síncrones: el punt d'inici/fi de la trama s'ubica segons el mateix criteri (màxim, mínim o pas per zero) del senyal de veu dins la seva periodicitat. Aquestes trames s'enfinestren amb finestres de Hamming, Hanning o derivats i llavors s'empren per crear noves periodicitats repetint-les o variant la distància entre elles quan és necessari per així poder satisfer l'especificació desitjada (Moulines i Charpentier, 1990; Moulines i Verhelst, 1995).

6. ***Multiband resynthesis OverLap and Add* (MBROLA)**: és una tècnica síncrona de superposició i suma que realitza la detecció de *pitch* automàticament, basada en models sinusoïdals per descompondre cada trama i a partir d'aquí resintetitzar a *pitch* i fase constant, alleugerant així molt problemes derivats d'una detecció de periodicitat errònea (Dutoit, 1997).

7. **Síntesi per coeficients Mel-freqüencials cepstrals (MFCC)**: realitza la síntesi a partir d'una representació basada en un modelat estadístic. En aquest cas, no és possible obtenir una síntesi completament refinada a partir del model, però sí que es pot fer una reconstrucció prou completa del tracte vocal. Les primeres tècniques feien servir un mètode d'excitació d'impulsos/soroll, mentre que les tècniques posteriors ja realitzaven una parametrització de la font (Koishida *et al.*, 1995; Stylianou, 2001; Chazan *et al.*, 2000).

Cal distingir de totes maneres que les tècniques de processament del senyal exposades intenten solucionar els diferents problemes derivats de la concatenació d'unitats. Aquests problemes es poden subdividir en problemes d'alt nivell (macroconcatenació) i problemes de baix nivell (microconcatenació) (Taylor, 2009). Els problemes de baix nivell apareixen a l'haver-hi discontinuïtats evidents en la forma d'ona del senyal que poden ser fruit de salts entre mostres consecutives generats en unir dos extrems: aquest fet es coneix com artefacte o *click*. Cal observar que un salt entre mostres actua de la mateixa manera que un impuls en alta freqüència i encara que això passi en l'àmbit d'una o dues mostres, a nivell auditiu pot durar suficientment per a que sigui perceptible per l'usuari fàcilment. En canvi, els problemes d'alt nivell són aquells que tenen a veure amb l'evolució del patró espectral en

una concatenació d'unitats. Per tant, la concatenació espectral ha d'esdevenir suavitzada o natural. Els problemes de discontinuïtat espectral es solucionen, la majoria de vegades, enfinestrant el senyal per trames en les unitats sonores o de manera global en les unitats sordes. L'enfinestrament provoca una suma del senyal del 50% de cada segment (*fading*) en el punt de concatenació per així assolir una transició suau (Harris, 1978; Nuttall, 1981).

Malgrat tot, les tècniques explicades no aconsegueixen arreglar del tot els errors de desalineament de fase en el punt de concatenació. Aquest desalineament normalment ve provocat per una detecció de la periodicitat errònia (marques de *pitch*) en la fase d'anàlisi del senyal. Per això és important disposar d'algorismes robustos a l'error per tal d'etiquetar i marcar les periodicitats de la bases de dades (Alías *et al.*, 2006*b*). L'error més típic en la detecció de periodicitats en el senyal és marcar el senyal per falsos màxims, falsos mínims o falsos passos per zero en el senyal (Gerhard, 2003). Una manera d'arreglar aquest problema és moure les marques d'inici i fi de les trames a concatenar fins a minimitzar la correlació creuada entre les dues trames (Stylianou i Syrdal, 2001).

Una altra característica a tenir en compte en el procés de disseny del mòdul de processament del senyal és el punt de la unitat on s'ha de fer la concatenació més enllà de la tipologia i mida d'aquesta. A nivell teòric es tria el centre de l'al·lòfon (en cas de difonema, l'extrem del difonema) com a punt de concatenació ja que és on hi ha la part del senyal que presenta menys variació, la més estable. Tot i així, hi ha tècniques de selecció del punt de concatenació que van més enllà de la teoria i trien el punt de concatenació fent una anàlisi acústica del senyal per tal de determinar el punt òptim de concatenació (Taylor *et al.*, 1991; Conkie i Isard, 1996).

La qualitat de les tècniques de concatenació explicades, sigui amb modificació prosòdica o no, és generalment superior respecte la generació de senyal segons predicció lineal per excitació d'impulsos. Les tècniques concatenatives tenen més o menys la mateixa qualitat entre elles, així que majoritàriament l'elecció d'una tècnica o d'altra es fa segons altres criteris com el temps de computació, la mida de les dades emmagatzemades o la robustesa dels mètodes d'anàlisi del senyal de per una veu determinada (Taylor, 2009).

Malgrat la semblança en la qualitat de les diferents tècniques, s'ha de considerar que mentre els models sinusoïdals assumeixen un model gairebé sense error en els senyals periòdics, els models per predicció lineal assumeixen un model menys robust i més sensible a l'error. De tota manera, els errors d'aquests models són relatius al fer sevir també parts residuals del senyal completament naturals. De tota manera, les tècniques que treballen directament sobre les mostres i per tant no assumeixen cap model del senyal (p.ex. TD-PSOLA), són més sensibles als errors dels algorismes marcadors de periodicitat, a di-

ferència del que succeeix en els models sinusoïdals o MBROLA. De fet, és per aquest motiu que es solen triar les tècniques que modelen el senyal, ni que sigui mínimament, en comptes de les que concatenen el senyal directament (Taylor, 2009). Tanmateix, el fet de modelar el senyal, tal com s'ha esmentat abans, assumeix l' obtenció d'una veu robotitzada, apecte crític en entorns on es vol una alta qualitat sintètica.

Donat el fet que la selecció d'unitats ha de ser robusta també als errors de marcatge per tal d'oferir una òptima qualitat, i seguint l'històric del grup de recerca (Alías i Iriondo, 2002) s'ha escollit la tècnica de TD-PSOLA sense modificació prosòdica per aquesta dissertació doctoral. El motiu de l'elecció ve donat per ser l'única tècnica que no modela el senyal explícitament i per tant és l'unica que pot fer imperceptible la robotització, proporcionant un millor escenari per la recerca dins la selecció d'unitats i alta qualitat.

2.3 Fase de disseny

Després d'exposar a grans trets com funciona l'estructura d'els CTP-SU, ja es poden estudiar totes les fases del seu disseny, des del tipus de la seva unitat acústica fins a la seva implementació final, passant per les aproximacions que hi ha tant com en la tria d'atributs utilitzats per la cerca en el corpus, la seva normalització, integració i ponderació final.

2.3.1 Unitats del corpus: informació acústica de mínima significança

El primer pas en el disseny d'un CTP-SU es determinar quina és la unitat acústica bàsica. És a dir, quina és la part mínima d'informació acústica que dota d'entitat un fragment de veu enregistrat i alhora es pot combinar fàcilment amb altres parts. En els sistemes de síntesi de segona generació normalment aquesta unitat era el difonema (apartat 2.1.2). Des d'una perspectiva històrica, en aquesta etapa del disseny s'han emprat tot tipus d'unitats mínimes. El recull realitzat a (Taylor, 2009) cita cada tipus pel seu nom més comú com es detalla a continuació:

- **Trames (*frames*)**: Trames de veu individuals de mida no definida, que poden ser combinades en qualsevol ordre (Hirai i Tenpaku, 2004).

- **Estats (*states*)**: Parts dels al·lòfons, normalment determinades per l'alineació dels estats d'un model ocult de Markov (Donovan i Woodland, 1999; Donovan i Eide, 1998).

- **Semifonemes (*half-phones*)**: El semifonema, més concretament el semial·lòfon, és la part del senyal de veu d'un al·lòfon delimitada per una part estable a un extrem i una part inestable (zona de concatenació) a l'altre extrem. El semifonema s'obté en tallar l'al·lòfon per la part estable quedant dos semifonemes a banda i banda de l'al·lòfon. Si tenim N al·lòfons diferents en un corpus, a la vegada existeixen $\approx 2N$ tipus diferents de semifonemes (Möhler i Conkie, 1998).

- **Difonemes (*diphones*)**: Tal com s'ha explicat a l'apartat 2.1.2, el difonema és aquell tram de senyal delimitat per les parts estables (centrals) de dos al·lòfons consecutius. En altres paraules, el difonema el conforma la transició des del centre d'un al·lòfon fins al centre del següent. Així, la paraula "hola" en català està conformada pels difonemes següents: /SIL-O/ /O-l/ /l-@/ /@-SIL/. N'hi ha com a màxim N^2 tot i que mai s'hi arriba degut a les regles fonotàctiques (veure apartat 2.1.2) (Coorman *et al.*, 2000; Clark *et al.*, 2004).

- **Al·lòfons (*phones*)**: Fonemes o al·lòfons en el seu sentit més estricte. N'hi ha N (Hunt i Black, 1996; Taylor i Black, 1998; Saito *et al.*, 1996).

- **Demisíl·labes (*demi-syllables*)**: És l'equivalent al semifonema però aplicat a la síl·laba: la demisíl·laba s'obté al tallar la síl·laba per la part estable de la seva vocal central quedant una demisíl·laba a banda i banda de la vocal. Si tenim M síl·labes diferents en un idioma tenim llavors $2M$ demisíl·labes (Pearson *et al.*, 1998).

- **Disíl·labes (*di-syllables*)**: Anàlogament al difonema, la disíl·laba és aquell tram de senyal de veu delimitat per les vocals de dos síl·labes consecutives. Hi ha M^2 disíl·labes (Chen, 2003; Law i Lee, 2000).

- **Síl·labes (*syllables*)**: Síl·labes tal com es defineixen normalment (Matousek *et al.*, 2005; Saito *et al.*, 1996; Yu i Wang, 2004).

- **Paraules (*words*)** Paraules tal com es defineixen normalment (Vosnidis i Digalakis, 2001; Stöber *et al.*, 1999; Portele *et al.*, 1996).

- **Frases (*phrases*)**: Frase considerada com a grups d'entonació no com a oració (Donovan *et al.*, 1999). Per una definició més concreta del concepte de frase veure l'apartat 2.2.1.

Normalment, la consideració principal per l'elecció de la unitat bàsica és la fonologia del propi idioma juntament amb les restriccions pròpies que puguin imposar altres mòduls del CTP-SU. Els idiomes europeus occidentals normalment consideren una unitat bàsica

basada en els derivats del fonema (al·lòfon, semifonema o difonema). En canvi, els idiomes orientals (xinès per exemple) consideren unitats que es basen en la síl·laba (síl·laba, demisíl·laba o disíl·laba - Taylor (2009)).

També s'ha d'observar que hi ha sistemes de CTP-SU que no queden limitats a un únic tipus d'unitat fixa sinó que es basen en un disseny més aviat heterogeni, que combina diferents tipologies d'unitat en funció de la seva disponibilitat en el corpus enregistrat (Yi, 2003; Bulyko, 2002). Aquests sistemes fan servir transductors d'estats finits ponderats (*Weighted Finite State Transducers* - WFST) per determinar a quin nivell d'unitat s'ha de realitzar la cerca. Llavors, les unitats candidates passen a ser des de frases fins a fonemes passant per tots els seus punts intermedis. Altres exemples els podem trobar en sistemes basats en frases portadores que contenen les construccions lingüístiques més típiques de l'idioma per després canviar només els noms o paraules clau (Donovan *et al.*, 1999). Cal afegir que a (Sagisaka *et al.*, 1992) s'indica que la clau per obtenir una veu sintètica d'alta qualitat passa per a obtenir seqüències llargues de parla contínua, i així es senten les bases de la selecció d'unitats no uniforme.

En el transcurs dels anys i gràcies a la recerca realitzada, les eines d'etiquetat i l'anàlisi automàtica de la parla han esdevingut més robustes i menys sensibles als errors. A mesura que s'ha anat avançant en els sistemes de síntesi de tercera generació, ha esdevingut fonamental el fet de concatenar unitats més llargues que el difonema en les seves diferents variants prosòdiques (Campbell i Black, 1997). A (Balestri *et al.*, 1999) s'exposa el fet que reduir les modificacions del senyal i el nombre de concatenacions d'unitats ha de ser l'objectiu principal pels sistemes de síntesi de tercera generació i per tant, els ajustos prosòdics s'han d'aplicar si i només si són absolutament necessaris.

És obvi, per tant, que la funció de selecció d'unitats en aquests sistemes de síntesi ha de tenir en compte les consideracions esmentades en el paràgraf anterior amb independència de la unitat bàsica considerada. Encara que la unitat sigui de baix ordre (al·lòfon, semifonema o difonema) sempre s'ha d'intentar trobar seqüències llargues prèviament enregistrades. Per tant, l'existència d'unitats mínimes ha de ser una millora en la flexibilitat de la síntesi però en cap cas en detriment de la qualitat de components ja enregistrades en el corpus.

2.3.2 Parametrització d'unitats: informació acústica i lingüística

Els paràmetres que es consideren dins del càlculs dels subcostos (SC) la funció de cost aporten informació dels diferents nivells de parametrització del corpus. La seva definició

(Rennison, 1994) constitueix una de les moltes línies d'investigació respecte els sistemes de tercera generació basats en síntesi concatenativa (Black, 2002). Tal com s'ha descrit en l'apartat anterior (2.3.1), hi ha moltes variants sobre les mesures a emprar, des de considerar només la informació acústica (Breen i Jackson, 1998; Black i Taylor, 1997*b*), estructures fonològiques només amb informació lingüística (Taylor i Black, 1999; Clark *et al.*, 2005) fins a sistemes híbrids (Yi, 2003; Bulyko, 2002). Aquests últims combinen informació lingüística per triar les unitats candidates amb informació acústica del punt de concatenació per a fer una concatenació òptima. Altres opcions inclouen agrupar la predicció prosòdica amb la selecció d'unitats i així modelar els subcostos mitjançant tècniques d'aprenentatge artificial segons la seva disponibilitat en el corpus (Campillo i Rodríguez-Banga, 2006).

Aquests subcostos es poden classificar bé respecte el tipus de domini que prenen els seus valors (Coorman *et al.*, 2000) o bé segons la informació que quantifiquen (Taylor, 2009).

Classificació basada en la representació

Segons la seva representació (Coorman *et al.*, 2000), es poden definir tres tipus de distàncies, en funció dels paràmetres que es vulgui comparar (Breuer i Abresch, 2004):

1. **Simbòlica**: determina el grau de semblança entre paràmetres categòrics de la unitat que són difícilment quantificables d'una manera contínua. Per exemple, en el cost d'*accent* de la sílaba, la distància es binaritza segons dos valors: 0, si totes dues unitats estan accentuades o no ho estan, o 1 si és que no coincideixen amb aquesta característica (Febrer, 2001; Campillo, 2005). No obstant, una mesura simbòlica no ha de ser binària, pot donar valors intermedis, com per exemple en el cas del difonema. Si un semifonema compleix la restricció d'accentuació de síl·laba i l'altre no el valor passa a ser 0.5. Així doncs, aquest tipus de distància està íntimament lligada al coneixement expert, en molts casos lingüístic, del paràmetre avaluat.

2. **Escalar**: són mesures que comparen paràmetres quantificables obtinguts de les unitats. Per exemple, la diferència en el *pitch* de les unitats, en la duració o en l'energia, entre d'altres. Es poden calcular com la diferència entre els paràmetres, absoluta, quadràtica, etc., o bé penalitzant la diferència en un cert sentit, per exemple, en la duració es pot penalitzar més que la unitat candidata presenti una duració menor a la desitjada. El treball de Febrer (2001) presenta una aportació per a la mesura de distàncies mitjançant funcions escalars contínues i discretes, definides per trams.

3. **Vectorial**: és una distància definida entre informacions generalment quantificables

de les unitats representades en forma de vectors multidimensionals. Per exemple, en el càlcul del C_C per a avaluar la continuïtat espectral se solen modelar les trames extremes de les unitats mitjançant paràmetres cepstrals (p.ex. vector de 24 enters) en l'escala Mel (*Mel Frequency Cepstrum Coefficients*), LSF (*Line Spectral Frequencies*) o (LSP *Line Spectral Pairs*), entre d'altres.

Classificació basada amb la parametrització

La modalitat més bàsica de representació d'una unitat és la seva forma d'ona. Aquesta té associada informació lingüística (o simbòlica) que almenys inclou la seva transcripció fonètica. En aquest sentit el conjunt de característiques no resulta trivial. La definició d'aquests paràmetres és el que diferencia els diferents CTP-SU. Principalment aquesta informació es classifica en informació acústica i informació lingüística.

1. **Informació acústica**: A partir de la forma d'ona, es pot derivar cap a qualsevol representació acústica: freqüència fonamental, energia, coeficients Mel-Cepstrum, etc. Normalment, aquesta informació es representa amb informació escalar o vectorial.
2. **Informació lingüística**: A partir del text es pot obtenir informació lingüística com ara: unitat prèvia, unitat posterior, posició en grup d'entonació, posició en el grup accentual, posició en la paraula, posició en la síl·laba, categoria gramatical, accentuació, etc.

2.3.3 Tria de paràmetres

A l'hora d'escollir els paràmetres que intervenen a la selecció d'unitats s'obre un ampli ventall de possibilitats. A priori, tot tipus d'informació resultaria vàlida: informació procedent de l'anàlisi lingüístic del text d'entrada o informació acústica (que pot venir tant de la predicció prosòdica com el propi etiquetat del corpus). Normalment aquestes característiques es separen en funció del tipus de distància: si aquesta analitza la diferència respecte una especificació prèvia s'anomenen subcostos de *target*. En canvi, si es mesura la discontinuïtat de certs paràmetres de la parla a través de la seva concatenació s'anomenen subcostos de concatenació. En la jerarquia següent s'exposa paràmetres típics en l'estat de l'art àmpliament acceptats (Taylor, 2009).

- **Subcostos de *target***
  - SUBCOSTOS LINGÜÍSTICS (DOMINI SIMBÒLIC)

1. *Subcost d'accentuació:* a cada al·lòfon se li assigna una etiqueta indicant si pertany a una síl·laba accentuada o no. Aquesta informació és tractada des dels nivells superiors de la jerarquia segons: no accentuat (0), accentuat parcialment (0.5 - en cas de tenir algun al·lòfon en síl·laba accentuada i l'altre no) i accentuat (1).
2. *Subcost de categoria gramatical (Part-Of-Speech):* La categoria gramatical de la paraula que conté la unitat bàsica considerada. Principalment s'usa per noms, verbs i adjectius. El seu subcost associat és 1 si no hi ha cap coincidència entre la unitat objectiu i la candidata, 0 si hi ha una correspondència exacte. En el cas de correspondències parcials, aquest cost sol ser ($\frac{CP}{TP}$) on CP (*Correspondant phones*) és el nombre total d'al·lòfons que corresponen amb l'especificació i TP (*Total Phones*) és el nombre total d'al·lòfons (Taylor i Black, 1998).
3. *Subcost de posició en grup d'entonació:* Posició de cada al·lòfon dins de la frase o grup d'entonació. El seu domini es defineix per inicial, mig, final i unitari.
4. *Subcost de posició en grup accentual:* Posició de cada al·lòfon dins del grup accentual. El grup accentual es conforma d'un conjunt d'al·lòfons delimitat per l'aparició de síl·labes tòniques. En català i en castellà, dins d'un grup accentual l'entonació no pateix cap variació (micromelodia - veure apartat 2.2.2) en el seu pendent ja que en aquesta la única variació ve inferida per la mateixa síl·laba tònica (Estruch *et al.*, 1996). D'una manera semblant al cas anterior, el seu domini queda definit per inicial, mig i final.
5. *Subcost de posició en paraula:* Posició de cada al·lòfon dins la paraula a la que pertany. El seu domini també queda definit per inicial, mig i final.
6. *Subcost de posició en síl·laba:* Posició de cada al·lòfon dins la síl·laba a la que pertany. El seu domini també queda definit per inicial, mig i final.
7. *Subcost de context:* Indica si les unitats anterior i posterior són les mateixes en les unitats candidates que en les unitats del text d'entrada. La seva funció de subcost queda definida d'una manera semblant a la que es fa servir al subcost de categoria gramatical.
8. *Subcost d'estil:* Indicació de l'estil d'entonació en la pronunciació del text. Clàssicament ve determinat per enunciatiu, interrogatiu o exclamatiu.
9. *Subcost d'èmfasi:* Indica si s'ha de realitzar èmfasi en alguna o altre paraula en la pronunciació del text. Pot venir marcada per la notació ToBI (Beckman i Hirschberg, 1994).

10. *Subcost d'emoció:* Marca la diferència d'emoció en la parla entre la unitat candidata i la unitat objectiu.

– SUBCOSTOS ACÚSTICS (DOMINI ESCALAR I VECTORIAL)

1. *Subcost de percepció del to (pitch):* Permet comparar la similitud de les freqüències fonamentals entre la unitat objectiu i la candidata, feta la mitjana per cada unitat.
2. *Subcost d'energia:* Codifica la similitud d'energia mitjana de la unitat candidata respecte a l'objectiu.
3. *Subcost de durada:* Determina la similitud entre les duracions dels al·lòfons de la unitat objectiu i la candidata.
4. *Subcosts de qualitat de veu (Voice-Quality*): La *Voice-Quality* fou definida per Trask (Trask, 1996) com la parametrització característica audible d'una veu individual, derivada d'una diversitat de característiques laríngiques i supralaríngiques que apareixen contínuament en la parla individual. Els tons dels sons de la parla que són naturals, distintius i són produïts per una persona en concret, defineixen una veu en particular. Segons el treball de Iriondo *et al.* (2008) aquests paràmetres serveixen per discriminar l'estil expressiu de locució d'una persona. Alguns exemples dels paràmetres de VoQ que es poden considerar com a càlcul d'aquests subcostos són:
   (a) *Jitter*: Mesura la variació cicle-a-cicle del període fonamental mitjanant la diferència de magnitud de dos períodes consecutius, dividit pel període mig.
   (b) *Shimmer:* Mesura la variació cicle-a-cicle de l'energia fent la mitjana la diferència d'energia de dos períodes consecutius, dividida per l'energia mitja.
   (c) *Glottal-to-noise excitation ratio (GNE)*: Quocient entre l'excitació de les vibracions de les cordes vocals i l'excitació degut a turbulències en forma de soroll (Michaelis *et al.*, 1997). A diferència d'altres paràmetres com el coeficient harmonic-soroll (HNR) o el soroll d'energia normalitat, GNE es gairebé independent del Jitter i el Shimmer.
   (d) *Hammarberg index*: Es defineix com la diferència de l'energia màxima entre les bandes de freqüència 0-2000Hz i 2000-5000Hz.
   (e) *Do1000*: És una regressió lineal (*least-squares*) del pendent espectral per sobre dels 1000Hz.

- **Subcostos de concatenació**

– SUBCOSTOS ACÚSTICS (DOMINI ESCALAR I VECTORIAL)

1. *Subcost de percepció del to (pitch):* Analitza la similitud de les freqüències fonamentals de les unitats en el punt de concatenació.
2. *Subcost d'energia:* Codifica la diferència de nivell energètic de les unitats a concatenar en el punt de concatenació.
3. *Subcostos basats parametrització cepstral (cepstrum):* Determina com n'és de bona la unió entre les unitats a nivell cepstral. El cepstre es defineix com la transformada discreta inversa de Fourier del logaritme de les magnituds d'un espectre. L'anàlisi es pot escalar segons l'escala Mel (Koishida *et al.*, 1995), que serveix per ponderar les bandes freqüencials segons la resposta auditiva humana. Normalment s'utilitzen 12 coeficients cepstrals més les seves derivades calculades dins d'una finestra de 20ms ubicada en el punt de concatenació (Campbell i Black, 1997). Hi ha diversos tipus d'aquests subcostos, on els més destacables són:
   (a) Subcost basat en coeficients cepstrals: és la discontinuïtat cepstral en el punt de concatenació calculant els coeficients cepstrals sense cap pas adicional ni tampoc cap simplificació del procés.
   (b) Subcost basat en coeficients cepstrals en l'escala Mel (MFCC): la parametrització cepstral a escala Mel, fa l'anàlisi en components de l'espectre freqüencial considerant la resposta freqüencial de l'aparell auditiu humà. Així, en comptes de donar igual importància a totes les bandes de freqüència de l'espectre, pondera aquelles freqüències en les quals l'oïda humana té una major perceptibilitat.
   (c) Subcost basat en coeficients cepstrals basats en predicció lineal (LPCC): Per motius d'optimització (p.ex. no calcular les covariàncies) resulta computacionalment més eficient fer la transformació a coeficients cepstrals a partir dels coeficients de predicció lineal (LPC). En aquest cas s'assumeixen en el càlcul els errors de precisió provocats pel filtre de predicció lineal previ.
4. *Subcosts basats en formants:* Els formants es defineixen com els $\mathcal{N}$ primers màxims d'energia en l'espectre d'un senyal (normalment $\mathcal{N}$ val 2, en alguns casos 3). El valor dels primers dos formants és el que ens permet reconèixer i parametritzar, per exemple, el triangle vocàlic; on el primer formant normalment determina l'obertura de la vocal (oberta tancada) i el segon formant determina el punt d'articulació de la mateixa (frontal, mig o

posterior).

5. *Subcosts basats en parametrització espectral (LSF):* Determina com n'és de bona la unió espectral de dues unitats en la seva representació com freqüències d'espectre lineal. Les freqüències d'espectre lineal (LSF) (Itakura, 1975; Soong i Juang, 1984) o també anomenades parelles d'espectre lineal (LSP) son una parametrització del senyal en el domini freqüencial semblant a l'anàlisi per codi de predicció lineal (LPC) (Kang i Fransen, 1987). A diferència dels LPCs, adopten característiques de l'error espectral en freqüència d'una manera selectiva, cosa que permet una quantificació més ajustada amb la percepció humana. A més, la facilitat d'estimar la sensitivitat a l'error espectral a cada línia de l'espectre permet codificar cada línia de l'espectre d'una manera eficient.
6. *Subcosts basats en codificació de predicció lineal (LPC):*
   (a) Predicció lineal perceptiva: d'una manera semblant als MFCC els coeficients del codi de predicció lineal originals de la veu s'adapten a la resposta auditiva de la manera següent: en comptes de calcular la funció d'autocorrelació en el domini temporal, es calcula la DFT de la finestra d'anàlisi, s'eleven els seus valors al quadrat i es calcula la DFT inversa. D'aquesta manera s'assegura la utilització de més pols per les freqüències baixes que per les freqüències altes amb un escalat semblant al dels MFCC o a l'escala Bark (Klabbers i Veldhuis, 1998).
   (b) Quocients (o ratios) d'àrea logarítmica de predicció lineal: originalment els *log area ratios* serveixen per representar els coeficients de reflexió. Aquests coeficients quantifiquen la incidència d'un model espectral en el seu coeficient posterior (semblant als LPC). Malgrat que els *log area ratios* són menys precisos (LSF) a l'hora de mesurar la transmissió per un canal, són més ràpids i senzills de calcular.

Es pot pensar (Taylor, 2009) que l'ajust dels pesos de la funció de cost ve determinat per una addició exhaustiva de tots els subcostos possibles, en comptes fer un procés de disseny acurat amb l'ajuda de coneixement expert. El procés de disseny s'entén com aquell procés que pondera les característiques més importants per guiar la selecció d'unitats. La importància de les característiques, tant si són inferides directament a partir del text d'entrada com les que són derivades a través d'un model entrenat, és relativa en funció de l'especificitat dels altres mòduls del CTP-SU (etiquetat del corpus, mida de la unitat, anàlisi lingüístic, predicció de la prosòdia, etc.). Massa sovint en el procés de disseny del CTP

es veuen les característiques com allò que s'ajusta aïlladament: primerament es defineixen els paràmetres i després s'incorporen dins la selecció d'unitats. Aquesta aproximació obvia un anàlisi en profunditat del que realment es necessita, derivant cap a un sistema poc eficient. A part, en utilitzar característiques a partir de l'anàlisi del text, s'ha de decidir si simplement s'usen les categoritzacions de l'anàlisi lingüístic o es processen d'alguna manera, per exemple normalitzant-les.

2.3.4 Formulació de característiques independents vs. formulació en l'espai acústic

Fins ara s'han explicat dues aproximacions sobre els subcostos de *target* que es poden categoritzar bé segons una formulació de característiques lingüístiques independents (*Independent Target Formulation* - IFF) o bé una formulació en l'espai acústic (*Acoustic Space Formulation* - ASF) (Taylor, 2009; Steiner *et al.*, 2010). La formulació per característiques independents ve determinada només per una anàlisi lingüística del text d'entrada, en canvi, la formulació en l'espai acústic ve determinada per una funció de predicció que mapa l'espai lingüístic a un espai acústic segons un model determinat de prosòdia.

Quan els paràmetres inferits directament del text d'entrada (paràmetres lingüístics) resulten insuficients per guiar la selecció d'unitats s'anomena l'especificació d'entrada com a ambigua. A l'aparèixer aquesta ambigüitat, la responsabilitat de la qualitat de la síntesi dins el CTP-SU recau en els mòduls posteriors (p.ex. composició de la forma d'ona, modificació del senyal). D'altra banda, si es defineix una especificació d'entrada basada en l'espai acústic (ASF) s'aconsegueix disminuir aquesta ambigüitat però a expenses d'introduir errors en aquesta predicció (p.ex. errors en l'etiquetat acústic del corpus); els processos que construeixen l'especificació no només introdueixen l'error propi d'aprenentatge sinó que propaguen els errors previs comesos en l'etiquetat del corpus. No obstant, si s'obtingués l'especificació acústica (prosòdica) a partir del text d'entrada amb la mateixa precisió que s'obté la informació lingüística, seria més fàcil assolir cerques d'unitats perfectes segons una especificació acurada de l'entrada.

Llavors, el disseny del CTP-SU comporta realitzar un judici previ abans d'establir el criteri a seguir en el mòdul de selecció d'unitats. En els pols oposats d'aquesta decisió s'hi troben per una banda l'opció d'ignorar la parametrització acústica de les unitats i per tant guiar la cerca només amb una parametrització lingüística (IFF) (Chu i Peng, 2001; Clark *et al.*, 2007) o bé, contràriament, realitzar una predicció prosòdica acurada (assumint els errors de predicció com a despreciables) i guiar la síntesi a nivell acústic sense infor-

mació lingüística (ASF) (Park *et al.*, 2003; Lee *et al.*, 2003). Aquesta última considera que implícitament s'ha inferit la informació lingüística en els valors acústics de la prosòdia. Altres enfocaments consideren treballar amb un híbrid considerant alhora la parametrització acústica i la parametrització lingüística, ponderant quina és més important en funció de la tipologia de cada unitat (Campillo *et al.*, 2005; Campillo i Rodríguez-Banga, 2006; Bonafonte *et al.*, 2008).

Des d'una perspectiva de detall de la funció cost de *target*, quanta menys ambigüitat hi hagi en l'especificació prèvia més precisa podrà ser la cerca d'unitats. Per exemple, no és el mateix realitzar una cerca a la base de dades en termes d'accentuació (*stress*) que en termes d'accentuació, entonació, frasificació, emoció o estil de la parla. De totes maneres, s'ha d'assolir un equilibri entre precisió i dispersió en les dades que conformen l'especificació original. Només si es disposa d'un nombre petit d'objectius de cerca prou diferents en l'especificació prèvia, s'obtindran cerques completes plenament satisfactòries, en canvi si hi ha un sistema de parametrització complet i exhaustiu rarament s'obtindran cerques satisfactòries del que es busca, depenent també del volum de veu que s'hagi enregistrat en el corpus. El nombre i tipus dels objectius no afecten a l'algorisme de selecció d'unitats en sí mateix, però tanmateix, disposar d'un ventall ampli de característiques en el vector d'especificacions obliga a disposar d'una funció de selecció més acurada i intel·ligent ja que aquesta ha d'establir prioritats entre els objectius parcials.

Alguns investigadors (Clark *et al.*, 2005; Colotte i Beaufort, 2005; Tihelka, 2005), que sostenen que la predicció prosòdica és senzillament transformar la representació lingüística original en l'espai acústic, conclouen que a priori no existeix cap motiu que impedeixi emprar únicament la parametrització lingüística d'entrada i reduir la complexitat del procés. El seu raonament ve donat per la hipòtesi, ja comentada, que aquesta transformació afegeix un error de precisió per culpa d'amitjanats, dades d'entrenament mal etiquetades, etc.

Llavors, l'elecció d'un tipus d'aproximació o altra esdevé un compromís entre la reducció de la dimensionalitat (menys característiques i menys complexitat del procés) i la precisió (disminució de l'error) en la cerca de les unitats. Els autors que defensen l'especificació simbòlica (lingüística) donen més importància en tenir poca dimensionalitat de manera precisa que tenir molts subcostos amb una especificació poc precisa i subjecte a errors. Posen l'exemple que a l'hora de predir un conjunt prosòdic aquest s'ha entrenat a partir d'una parametrització lingüística d'alt nivell, però després aquesta informació és descartada assumint que la única inferència de la informació lingüística en la parla només és reflexada en el conjunt prosòdic esmentat (clàssicament entonació, energia i ritme). Aquesta assumpció

és bastant restrictiva ja que hi ha experts que demostren que la informació lingüística i expressiva d'alt nivell també pot influir en paràmetres de la qualitat de veu (*voice-quality*) i altres efectes espectrals (Monzo *et al.*, 2007).

La tria de característiques depèn de molts factors i no es pot dir si una aproximació és millor a una altra. Tanmateix els treballs publicats els últims anys han passat d'utilitzar informació acústica de baix nivell a informació lingüística d'alt nivell, per exemple (Tihelka, 2005; Colotte i Beaufort, 2005). Llavors, sembla que la tendència és que els sistemes de síntesi han de considerar les característiques lingüístiques d'alt nivell i llavors entrenar la importància de cada subcost *target* mitjançant algun algorisme d'aprenentatge automàtic.

En tots dos casos, es pot usar el principi d'espai perceptiu (Strom i King, 2008). Segons aquest principi les diferents combinacions de característiques prosòdiques queden representades en l'espai perceptiu que és n-dimensional. Llavors, les unitats amb vectors de característiques prosòdiques similars, sigui quina sigui la seva formulació, queden agrupades dins la mateixa regió en l'espai perceptiu. Un cop definit l'espai perceptiu, s'obté la distància entre dos unitats mitjançant una mètrica dins el mateix espai, per exemple, una distància euclidiana. Segons aquest principi, l'equació clàssica de Hunt i Black per definir la funció de selecció d'unitats (veure equació 2.4) quedaria representada en l'espai perceptiu com una distància de Manhattan ponderada (eq. 2.1) (Hunt i Black, 1996).

De tota manera, aquesta discussió no està ni molt menys tancada i la seva resolució és un dels aspectes que pretén discutir aquesta tesi doctoral d'una manera empírica i basada en el modelat de la percepció humana usant algorismes evolutius.

2.3.5 Integració dels subcostos

En general els subcostos de target i concatenació s'integren en una sola funció de cost global seguint una distància de Manhattan ponderada (eq. 2.4), o aplicant alguna altra mètrica, com les descrites a (Toda, 2003; Toda *et al.*, 2004, 2006), on a més d'estudiar el subcost més adequat per a cada tipus de paràmetre, també s'analitza la millor manera, en termes de correlació subjectiva, d'integrar aquests subcostos en la funció de cost. En aquest cas, proves perceptives aïllades fetes en grup d'ajust determinen la distància (Manhattan, euclídea, etc.) i els pesos de cada subcost.

Normalització de subcostos i reducció de la dimensionalitat Integrar les distàncies associades a cada característica directament comporta un biaix cap aquella distància amb un marge de valors més gran. En aquest sentit resulta apropiat preprocessar les distàncies i

transformar-les perque la seva integració resulti més equitativa. En aquest sentit, també és important que els valors de les distàncies adoptin funcions de densitat semblants en termes d'asimetria i curtosi, i alhora semblants a la distribució normal (Tukey, 1957).

En l'apartat 2.3.3 s'ha explicat que l'excés de dimensionalitat en la funció de cost suposa augmentar el problema de la seva posterior ponderació i integració (Hunt i Black, 1996). Aquesta integració de subcostos de diferent naturalesa es coneix com a procés de linealització. Si es pot fer una anàlisi previa de les dades tenint en compte les dependències dels subcostos entre ells (p.ex. correlacions lineals) es pot reduir la dimensionalitat i alhora obtenir unes dades normalitzades dins un interval $[0, 1]$. A tal efecte, es poden fer servir algorismes de preprocessament de dades basats en diagonalització de matrius, canvi de base i reducció de la dimensionalitat tals com l'anàlisis semàntic latent (LSA) (Bellegarda, 2009).

Integració ponderada de subcostos Després de la normalització, l'últim pas per obtenir la funció de cost és la ponderació dels diferents subcostos, per així obtenir un sol cost que determini la indoneïtat d'una unitat específica dins la seqüència d'unitats. Tal com s'ha explicat a l'apartat 2.3.4, a través d'un combinació lineal ponderada, s'obté la distància entre dues unitats combinant diferents objectius. Tanmateix, l'optimització lineal indicada per Hunt i Black (1996) és un cas concret d'integració mitjançant una distància de Manhattan ponderada.

Si es representen els diferents subcostos en un espai de subcostos n-dimensional, aquest espai es pot escalar i rotar per tal que representi més bé la diferència de les unitats segons un criteri extern (p.ex. perceptiu). En el cas perceptiu, s'obté un espai d'optimització nou rotant les dades a través d'una matriu de pesos, que prioritzen els subcostos més sensibles a la percepció de l'oïda davant aquells els subcostos imperceptibles auditivament (Chu i Peng, 2001; Toda *et al.*, 2006).

La ponderació dels esmentats pesos és una de les línies de recerca que segueix sense tancar-se en la selecció d'unitats. A l'apartat 2.4 s'estudien en profunditat les diferents metodologies existents per l'ajust d'aquests pesos, tant des d'un punt de vista objectiu, sense intervenció humana, com des d'un punt de vista subjectiu, amb supervisió humana durant el procés.

2.4 Ajust objectiu de pesos vs. ajust perceptiu de pesos

L'ajust dels pesos per a la funció cost clàssica és una de les línies de recerca no tancades en CTP-SU, tal com s'ha esmentat en l'apartat 2.3.5. La dificultat principal es troba en desenvolupar un procés genèric que automàticament ajusti els pesos segons les necessitats de cada idioma, estil i enregistrament, per així obtenir parla sintètica d'alta qualitat. Contràriament a altres aproximacions a la síntesi de la parla (LPC o HMM), històricament l'ajust dels pesos s'ha realitzat de manera molt artesanal, o bé emprant tècniques automàtiques que assumien principis no contrastats per determinar la qualitat (p.ex., distàncies cepstrals).

Les aproximacions de l'ajust de pesos poden ser divididas en tres grups: *i*) ajustos a mà (Coorman *et al.*, 2000), *ii*) mètodes computacionals purament objectius (Hunt i Black, 1996; Meron i Hirose, 1999; Park *et al.*, 2003; Alías i Llorà, 2003) i *iii*) tècniques d'optimització perceptiva (Lee *et al.*, 2003; Peng *et al.*, 2002; Toda *et al.*, 2006). A continuació es detallen els grups en profunditat.

2.4.1 Ajust a mà

Una primera manera d'obtenir els pesos de la funció de cost es basa en ajustar-los a mà a través d'un procés de prova i error (Breen i Jackson, 1998; Clark *et al.*, 2004; Coorman *et al.*, 2000; Clark *et al.*, 2007). El procés guarda certa analogia amb una equalització acústica. No obstant, l'optimització s'esbiaixa cap el coneixement expert de la persona que realitzen l'ajust. Normalment, en l'etapa de disseny s'escolten resultats preliminars de la selecció d'unitats i tot seguit, es passa a "equalitzar" els pesos ajustant-los mitjançant prova i error, en base a un criteri fruit del compromís entre el coneixement previ i la veu escoltada.

Aquest enfocament, tot i que pot produir uns resultats acceptables, té una part artística i personal que no permet establir cap metodologia general d'ajust de pesos més enllà de quatre punts d'assessorament. Tampoc permet establir un criteri unívoc quan aquest ajust es fa a través de diferents persones: En certa manera resulta contradictori anomenar "metodologia" un ajust manual quan realment el que succeeix és que no es pot oferir una metodologia exacta d'ajust. No obstant això, l'ajust manual obté bons resultats (Clark *et al.*, 2007; Schröder *et al.*, 2009) avui en dia. Malgrat això, cal destacar que aquests treballs, sistemes típics dins CTP-SU, disposen d'entre 10 i 20 pesos, quantitat assumible per a una persona.

De tota manera, el clar avantatge de fixar els pesos a mà és que, mitjançant l'ús de

judicis globals basats en coneixement expert, es prioritzen els subcostos més importants per obtenir una bona qualitat sintètica de manera explícita. Així s'evita l'assumpció dels mètodes d'optimització automàtica que consideren igual importància en els errors (p.ex., els errors de concatenació /r/ són igual d'importants que els errors de concatenació en les /a/). Explorant les propietats en les dades, es pot veure que no tots els errors en la selecció d'unitats són igualment assumibles o mereixedors de ser descartats. En certa manera, s'hauria d'establir una gradació de l'error ja que no és el mateix seleccionar unitats on la seva bondat ve determinada per les unitats veïnes, que seleccionar unitats mal etiquetades o amb artefactes que són dolentes en qualsevol context. Una supervisió manual permet tenir una visió global i per tant que l'ajust sigui satisfactori en aquest sentit.

Aquesta superioritat permet concloure que, fins al moment, la percepció humana imposa una complexitat en la formulació actual de la selecció d'unitats que limita l'ús de tècniques d'optimització convencionals. Es pot dir que existeix un coneixement implícit que no s'aconsegueix representar matemàticament i per tant provoca restriccions al problema d'optimització segons una aproximació computacional clàssica. En aquest sentit, no es nega l'optimització basada en l'estadística ni l'aprenentatge automàtic: els exemples del seu bon funcionament són nombrosos en diferents àmbits fins i tot en el camp de la síntesi (Tokuda *et al.*, 2003). Simplement s'afirma que la mesura de qualitat que s'empra en aplicar aquests mètodes al disseny de la funció de cost resulta insuficient a hores d'ara.

2.4.2 Ajust de pesos objectiu

Els mètodes d'ajust de pesos purament objectius són aquells que realitzen l'optimització de la funció de cost sense participació humana i basant-se únicament amb funcions de distància objectives (en el domini espectral o temporal entre veu natural i veu sintètica). El fet de poder calcular els pesos de manera automàtica tantes vegades com sigui necessari permet, com s'exposarà més endavant, superar problemes crítics en l'ajust interactiu o perceptiu tals com són les contradiccions i l'ambigüitat en l'ajust de la funció de selecció. A més, a nivell de precisió permet calcular tantes combinacions de pesos com unitats hi hagin al corpus sense que resulti necessari assumir un vector de pesos per tot el corpus.

Weight Space Search

A (Hunt i Black, 1996; Campbell i Black, 1997) es presenta el primer mètode per a realitzar l'ajust de pesos de forma automàtica, anomenat WSS (*Weight Space Search*). El mètode es basa en una mesura acústica obtinguda mitjançant distàncies cepstrals entre senyal real

i senyal resintetitzat. L'algorisme realitza una exploració exhaustiva de diferents combinacions de pesos considerant cada versió enregistrada en el corpus com una especificació *target*.

El motiu d'elecció de les distàncies cepstrals ve donat pel treball de Rabiner i Juang (1993), que indica que la millor mesura que mapa la diferència auditiva perceptiva entre dos senyals és la distancia cepstral euclidiana entre vectors alineats en temps mitjançant *Dynamic Time Warping*. Posteriorment es demostrarà que aquesta mesura ha estat posada en dubte segons la investigació realitzada en posterioritat (apartat 2.4.3).

L'algorisme resintetitza frases ja existents en el corpus amb diferents conjunts de pesos, sense considerar les unitats de la pròpia frase que són extretes prèviament. Així s'obté un conjunt de N_{WSS} diferents frases sintetitzades amb M_{WSS} diferents combinacions de pesos. Per a cada posició de la matriu $M_{WSS} \times N_{WSS}$ es pot avaluar la bondat de la síntesi amb una distància objectiva respecte la frase natural disponible al corpus. A continuació s'especifica amb detall:

1. En primer lloc, s'esborra una frase del corpus, de forma que els seus segments no estan disponibles per a ser seleccionats. Els seus segments naturals s'utilitzen per a definir vector d'especificacions (*target*) de la frase per a assegurar que aquesta prova és independent d'altres mòduls del CTP-SU (p.ex. la predicció de la prosòdia).

2. L'algorisme de Viterbi troba la seqüència que millor minimitza la funció de cost ponderada.

3. Els vectors de coeficients cepstrals (*cepstrum*) de les unitats seleccionades s'alineen temporalment amb els coeficients dels segments de *target*, i es calcula la distància euclidiana mitjana cepstral entre les unitats del segment de *target* (l'original) i el seleccionat (el que el substitueix).

4. El procés es repeteix per tots els conjunts diferents de pesos i per diferents frases.

Finalment, es trien aquells valors de pesos que minimitzen la distància mitjana cepstral obtinguda de les diferents frases.

En la seva realització pràctica (Hunt i Black, 1996) van provar de 3 a 5 possibles valors de pesos tant per característiques de context prosòdic com fonètic. Totes les combinacions possibles d'aquests valors foren provades en un mínim de 10 frases d'entrenament que requerien la síntesi i la comparació d'unes 100.000 formes d'ona. Tanmateix, aquest mètode d'entrenament té limitacions importants. Fonamentalment, el temps d'execució creix exponencialment segons el número de pesos que s'estan entrenant i amb el número de valors

considerats per a cada pes. Aquest problema el van reduir tot entrenant els pesos en passades múltiples, però malgrat tot, encara es requerien més de 150 hores d'entrenament per a un corpus de veu d'unes 40.000 unitats (equivalent a aproximadament 1 hora de parla) en una màquina *Sun SPARCStation 20* del 1995.

Tot i així, els autors conclouen que la distància objectiva hauria d'incloure un biaix per incrementar la sensibilitat per a punts de concatenació local amb errors de *burst* i així aproximar millor les preferències humanes. Caldria notar també que aquestes mesures acústiques, degut a la necessitat de l'alineament temporal, són cegues a duracions inapropiades, i no mesuren la degradació d'obtenir durades excessivament curtes en les unitats. Per aquest motiu, és possible que no estiguin ben correlades amb les percepcions humanes.

En una posterior versió, aquest mètode s'optimitza mitjançant heurístiques i poda (Meron i Hirose, 1999) per reduir el temps d'execució.

Regressió multilineal

El segon mètode basat en distàncies cepstrals presentat a (Hunt i Black, 1996) realitza d'una regressió multilinial (*Multilinear Regression* - MLR). En aquest cas, les distàncies cepstrals són considerades com el coeficient d'ordre zero d'una regressió lineal amb els subcostos associats. Així, una regressió segons el mètode dels mínims quadrats obté coeficients d'una recta entre els subcostos i les distàncies. Un cop s'han obtingut aquests coeficients, que posteriorment es normalitzen per a que sumin 1, aquests són considerats com a pesos. Per a entendre millor el procés tot seguit s'explica algebraicament.

Aproximació general La idea és avaluar l'efecte de la substitució, dins el seu context, d'una unitat del corpus per les altres versions candidates. Principalment s'observa quines variacions presenten els subcostos en relació amb les distàncies cepstrals. Es pot considerar, per tant, un vector D de N_{VER} distàncies cepstrals associat a la matriu $N_{VER} \times (M_{COST} + 1)$ de subcostos SC. La matriu transposada de la matriu SC serà SC^T[2]. En aquesta matriu les files i es corresponen a les diferents versions candidates i les columnes corresponen als j subcostos. Llavors D_i representa la distància cepstral de la unitat original respecte la versió i, SC_{ij} el subcost segons el paràmetre j de l'esmentada versió i, i e_i l'error de predicció (valor residual) desconegut de la versió i. En notació matricial:

[2]En aquesta dissertació doctoral la transposició de matrius es detalla segons l'operador $(\cdot)^T$

$$D = \begin{bmatrix} D_1 \\ \vdots \\ \vdots \\ \vdots \\ D_{N_{VER}} \end{bmatrix}, \; SC = \begin{bmatrix} 1 & SC_{11} & \cdots & \cdots & \cdots & SC_{1M_{COST}} \\ 1 & \vdots & \ddots & \ddots & \ddots & \vdots \\ 1 & \vdots & \ddots & SC_{ij} & \ddots & \vdots \\ 1 & \vdots & \ddots & \ddots & \ddots & \vdots \\ 1 & SC_{N_{VER}1} & \cdots & \cdots & \cdots & SC_{N_{VER}M_{COST}} \end{bmatrix}, \; e = \begin{bmatrix} e_1 \\ \vdots \\ \vdots \\ \vdots \\ e_{N_{VER}} \end{bmatrix} \tag{2.7}$$

on M_{COST} és el nombre de pesos que es volen obtenir, un per a cada subcost i N_{VER} és el nombre de versions de la unitat. En notació matricial, el model de regressió multiple es pot expressar com:

$$D = SC \cdot W^T + e \tag{2.8}$$

on W (de l'anglès *weight*) és un vector de coeficients de regressió (on W_1 és la component contínua) i e és l'error residual. L'objectiu és minimitzar la suma dels valors residuals quadrats:

$$\min_W ||e||_2 \tag{2.9}$$

Els coeficients de regressió que satisfan aquest criteri es poden trobar mitjançant el sistema d'equacions lineals següent, obtingut multiplicant ambdós costats per SC^T:

$$SC^T \cdot D = SC^T \cdot SC \cdot W^T \tag{2.10}$$

A continuació es multipliquen tots dos costats de l'equació (2.10) per les equacions normals a través de la matriu inversa $SC^T \cdot SC$ per així obtenir W^T:

$$W^T = (SC^T \cdot SC)^{-1} SC^T \cdot D \tag{2.11}$$

Quan els subcostos SC són linealment independents (és a dir, quan $SC^T \cdot SC$ és de rang complet), només hi ha una única solució al sistema d'equacions lineals.

Una altra manera és solucionar directament el sistema de l'equació (2.10) amb LS (*Least Squares*) si aquest és indeterminat, és a dir quan $N_{VER} < M_{COST}$, o usant la factoritza-

ció QR per als sistemes sobredeterminats (p.ex. $N_{VER} > M_{COST}$). Aquest mètode és més general i no requereix tot el temps associat a invertir la matriu que necessita el mètode convencional de resolució. D'altra banda, els mètodes de descomposició en valors singulars també hi són aplicables, però normalment són significativament més elevats en cost computacional i només són avantatjosos quan hi ha una dependència lineal molt forta entre subcostos, que disminueix la qualitat dels models (Breuer i Abresch, 2004). La tercera manera de resoldre el problema de la dependència lineal entre variables (si el determinant de la matriu $SC^T \cdot SC$ és positiu) és utilitzant la inversió de matrius normal, sense calcular la matriu transposada. A més s'hi suma la restricció $w_j > 0 : \forall w_j \in W$ per la qual els pesos de la funció de cost no poden ser negatius (no té sentit pensar en subcostos negatius quan estem parlant de diferències absolutes ja normalitzades). A aquest efecte s'explica primer com resoldre un problema de mínims quadrats (LS) sense aquesta restricció i posteriorment es detalla l'algorisme NNLS (*Non-Negative Least Squares*) (Lawson i Hanson, 1995) per adaptar el problema a la restricció existent.

Mínims quadrats Un principi fonamental dels mètodes de mínims quadrats (LS), i en particular de la regressió lineal múltiple, és que la variància de les variables dependents pot ser dividida en parts, d'acord amb l'origen de les dades. Si es suposa que en una variable dependent (distàncies cepstrals) es modela (mitjançant regressió) a través d'un o més coeficients (pesos multiplicats als subcostos) i, per conveniència, la variable dependent s'escala de tal manera que la seva mitjana sigui zero, llavors s'obté (Netter *et al.*, 1990) que la suma quadràtica dels valors observats de la variable depenent (variància les distàncies cepstrals) s'assoleix amb la suma quadràtica dels valors predits (subcostos ponderats) més la suma quadràtica dels errors residuals conformant així l'equació de mínims quadrats:

$$\sum (d - \bar{d})^2 = \sum (\hat{d} - \bar{d})^2 + \sum (d - \hat{d})^2 \tag{2.12}$$

on el terme $\hat{d}$ es refereix al valor de la distància predita, el terme $\bar{d}$ és la mitjana de d, i $\sum (d - \bar{d})^2$ és la suma quadràtica de les desviacions dels valors observats respecte la mitjana de la variable dependent. La part dreta de l'equació consta de:

- SS_{Model} la suma quadràtica (*Squared Sum*) de les desviacions dels valors predits ($\hat{d}$) respecte la mitjana de la variable ($\bar{d}$).
- SS_{Error} la suma quadràtica (*Squared Sum*) de les desviacions dels valors observats (d) menys els valors predits ($\hat{d}$), és a dir, la suma quadràtica de l'error (residus).

Dit d'una altra manera,

$$SS_{Total} = SS_{Model} + SS_{Error} \tag{2.13}$$

Cal notar que SS_{Total} val el mateix per qualsevol conjunt de dades particular, però SS_{Model} i la SS_{Error} varien amb l'equació de regressió. Assumint de nou que la variable dependent està escalada de tal manera que la seva mitjana és zero, la SS_{Model} i el SS_{Error} es poden calcular de la forma següent:

$$SS_{Model} = W \cdot SC^T \cdot D \tag{2.14}$$

$$SS_{Error} = D^T \cdot D - W \cdot SC^T \cdot D \tag{2.15}$$

I assumint que $SC^T \cdot SC$ és una matriu de rang complet, s'obtenen els índexs de fiabilitat d'una regressió lineal:

$$\text{RMSE} = SS_{Error} \tag{2.16}$$

$$R^2 = 1 - \frac{SS_{Error}}{SS_{Total}} \tag{2.17}$$

$$s^2 = \frac{SS_{Error}}{N_{VER} - M_{COST} - 1} \tag{2.18}$$

$$F(m, n-m-1) = \frac{SS_{Model}}{M_{COST} \cdot s^2} \tag{2.19}$$

on RMSE és l'error mig quadràtic de la regressió, R^2 és el coeficient de correlació al quadrat, altrament anomenat coeficient de determinació que mesura qualitat del model de *fitness* en funció de la desviació (percentatge) de dades que aquest pot explicar, s^2 és l'estimació no esbiaixada del residu o de l'error de la variància, i F és el criteri de Fisher de $(M_{COST}, N_{VER} - M_{COST} - 1)$ graus de llibertat. Si $SC^T \cdot SC$ no és de rang complet, el $rang(SC^T \cdot SC) + 1$ es substitueix per m (Netter *et al.*, 1990).

Mínims quadrats no negatius Segons (Lawson i Hanson, 1995) el problema de mínims quadrats no negatius (*Non Negative Least Squares* - NNLS) es defineix per la condició 2.20.

$$\text{Minimitzar } ||SC \cdot W^T - D|| \text{ condicionada a } w \geq 0 : \forall w \in W \tag{2.20}$$

Aquest problema de NNLS es resol amb l'algorisme 2.2 extret i adaptat de Lawson i Hanson (1995).

Igual que un problema de mínims quadrats normal, l'algorisme parteix inicialment d'una matriu SC de $N_{VER} \times M_{COST}$ dimensions, on existeixen N_{VER} versions d'unitats i M_{COST} subcostos. El vector D adopta igual mida que les versions (N_{VER}). Durant l'execució de l'algorisme es defineixen i modifiquen els conjunts d'índexs $\mathcal{P}$ (positius) i $\mathcal{Z}$ (zeros). El valor de les variables indexades per $\mathcal{Z}$ és 0. En canvi, els valors de les variables indexades per $\mathcal{P}$ adoptaran qualsevol valor diferent de 0. Si alguna d'aquestes variables pren un valor que no sigui positiu, l'algorisme intentarà refer el valor de la variable a un valor positiu o altrament fixar el valor d'aquesta variable a 0 i moure el seu índex del conjunt $\mathcal{P}$ al conjunt $\mathcal{Z}$.

Un cop l'algorisme acaba, el vector W és la solució al problema i el vector e és el vector d'error acumulat per a cada relació pes-subcost (vector dual).

A l'acabar, el vector de solucions W satisfà per a tot $w_j \in W$

$$w_j > 0 \quad j \in \mathcal{P} \tag{2.21}$$
$$w_j = 0 \quad j \in \mathcal{Z} \tag{2.22}$$

i per tant és la solució pel problema de mínims quadrats

$$SC_{\mathcal{P}} \cdot W^T \cong D \tag{2.23}$$

El vector dual e satisfà

$$e_j \quad : \quad j \in \mathcal{P} \tag{2.24}$$
$$e_j \leq 0 \quad : \quad j \in \mathcal{Z} \tag{2.25}$$
$$e \quad = \quad SC^T(D - SC \cdot W^T) \tag{2.26}$$

A (Lawson i Hanson, 1995) es demostra que les equacions (2.21, 2.22, 2.24, 2.25 i 2.26) constitueixen les condicions (conegudes com a condicions de Kuhn-Tucker) que caracteritzen un vector de solucions W per un problema de mínims quadrats no negatius. L'equació (2.23) és una conseqüència de les equacions (2.22, 2.24 i 2.26).

Un cop vistos els fonaments matemàtics de la regressió multilineal, es pot concloure que és un mètode senzill per aconseguir una relació lineal entre els diferents subcostos i

Algorisme 2.2 Algorisme NNLS per resoldre el problema de mínims quadrats no negatius.

procedure NNLS (SC, D)

1: $N_{VER} = size\_row(SC)$
2: $M_{COST} = size\_col(SC)$
3: $\mathcal{P} := NULL$;
4: $\mathcal{Z} := \{1, 2, \ldots, M_{COST}\}$
5: $w := 0 : \forall w \in W$
6: $e := SC^T(D - SC \cdot W^T)$ {Calcular el vector M_{COST}-dimensional e}
7: **if** $((\mathcal{Z} = \emptyset) \cup (e_j \leq 0 : \forall j \in \mathcal{Z}))$ **then**
8: retorna W;
9: **end if**
10: $t := \{t \in \mathcal{Z} | e_t = \max\{e_j : j \in \mathcal{Z}\}\}$
11: Movem t de $\mathcal{Z}$ a $\mathcal{P}$
12: Sigui $SC_{\mathcal{P}}$ una matriu de $N_{VER} \times M_{COST}$ definida per:

$$\text{Columna } j \text{ de } SC_{\mathcal{P}} := \begin{cases} \text{columna } j \text{ de } SC & \text{si } j \in \mathcal{P} \\ 0 & \text{si } j \in \mathcal{Z} \end{cases}$$

Es calcula el vector M_{COST}-dimensional v com una solució normal del problema de mínims quadrats $SC_{\mathcal{P}} \cdot v^T \cong D$. Val a dir que aquest problema només determina els components de $v_j | j \in \mathcal{P}$ i per tant $v_j = 0 : \forall j \in \mathcal{Z}$
13: **if** $e_j > 0 : \forall j \in \mathcal{P}$ **then**
14: $W := v$
15: GOTO(6)
16: **end if**
17: $q := \{q \in \mathcal{P} | \frac{w_q}{w_q - v_q} = \min\{\frac{w_j}{w_j - v_j} : v_j \leq 0, j \in \mathcal{P}\}\}$
18: $\alpha := \frac{w_q}{w_q - v_q}$
19: $W := W + \alpha(v - W)$
20: Moure tots els índexs $j \in \mathcal{P}$ tals que $e_j = 0$ des de $\mathcal{P}$ a $\mathcal{Z}$
21: GOTO(8)

les distàncies cepstrals. Val a dir que el mètode només resultarà útil si la dependència entre les dades és lineal i no funcionarà per una dependència polinòmica.

Selecció no uniforme mitjançant arbres de decisió

Un altre mètode d'ajust automàtic de la funció de cost és el desenvolupat per Colotte i Beaufort (2005), on es presenta una selecció no uniforme basada en un agrupament previ d'unitats segons el seu comportament acústic. Un cop obtinguts els grups, s'ajusten els pesos que millor discriminen els grups mitjançant un arbre de decisió. El seu treball assumeix que els models prosòdics no permeten variabilitat en la predicció i per tant aquesta resulta poc natural. De fet, (Campillo i Rodríguez-Banga, 2006) també parteix d'aquesta mancança. Aquest fet implica que la concentració de l'oient minvi a mesura que avança la síntesi. Llavors, (Colotte i Beaufort, 2005) decideix emprar només subcostos lingüístics per evitar l'ús d'un sol model prosòdic. Posteriorment realitzen una ponderació dels pesos a nivell d'unitat en funció del guany d'entropia que aporta el subcost dins l'arbre. Aquesta aproximació permet no restringir les unitats seleccionades a una prosòdia predeterminada. A continuació es detalla la seva proposta:

El seu treball empra al·lòfons (en comptes de difonemes) i subcostos lingüístics o altrament anomenats de notació simbòlica (veure apartat 2.3.4). Per un al·lòfon concret agrupen les versions de les unitats en diferents grups (clústers) calculats en funció d'una mesura de similitud (primer en funció de la durada i després en funció de la distància Kullback-Leiber (Veldhuis i Klabbers, 2003)). Per tal d'obtenir aquests clústers es segueixen els següents passos: en primer lloc *i*) s'inicialitza l'algorisme *k-means* per obtenir una primera agrupació basada en durades. *ii*) Després es refinen els grups iterant el *k-means* en funció de la distància de Kullback-Leiber. L'algorisme de *k-means* determina els grups en funció la seva mitjana i desviació típica. En tercer lloc *iii*) es calcula el nombre òptim de subgrups per a cada grup de durada prèviament establert. En aquest cas, el nombre òptim de subgrups es calcula intentant obtenir la mateixa variància a cada subgrup (maximització de quocient de variàncies). Un cop obtinguts tots els subgrups es procedeix a *iv*) construir un arbre de decisió (*IGTree*) que determini la importància de les diferents característiques lingüístiques per identificar un subgrup ja existent. Val a dir que en aquest cas l'arbre de decisió no realitza el *clustering* pròpiament dit, sinó que serveix jerarquitzar les característiques lingüístiques per identificar un grup ja existent (obtingut d'una *clusterització* prèvia feta amb *k-means*). L'arbre de decisió es construeix mitjançant el guany d'informació (*Information Gain* - IG) o altrament dit entropia. Per tant, realitza una prioritització de les característiques lingüístiques d'acord amb la seva informació intrínseca en

els grups acústics: quan una característica adopta valors baixos d'entropia (més amunt en la jerarquia de l'arbre), més significant i pertinent és la característica a l'hora de classificar. Com a últim pas *v*), per a cada paràmetre lingüístic, es calcula el guany d'informació segons els grups construïts pel *k-means*. Aquest fet permet determinar les característiques de classificació entre tots els nivells de l'arbre de decisió (*IGTree*). A partir d'aquests valors s'obté, per a cada característica, el valor del pes pel calcul del cost de *target*. Concretament els pesos s'obtenen mitjançant una escala logarítmica del quocient de guany (*Gain Ratio* — *GR*) detallat en l'equació 2.27:

$$W_k^l = 2 \times \log(10 \times GR(k,l)) \tag{2.27}$$

on W_k^l és el pes del subcost k per l'al·lòfon l i GR és el el quocient del guany d'informació per la característica k dins dels grups de l'al·lòfon l.

2.4.3 Ajust perceptiu

Tal com s'ha comentat al subapartat 2.4.1, la millor qualitat de la síntesi s'obté amb intervenció humana i manual, la qual troba similituds entre unitats naturals i unitats sintètiques. Només segons un criteri humà es pot disposar d'informació fiable en el disseny de la funció de cost. Aquest fet és degut a la debilitat de l'enfocament d'ajust purament objectiu, al dependre d'una funció de distància incompleta, que només en alguns aspectes correspon a la percepció humana (Campbell i Black, 1997). Els defectes amb detall s'han explicat a l'apartat 2.4.2.

Distàncies basades en el model auditiu

Aquesta aproximació intenta construir una nova mesura de degradació de la naturalitat basada basada en les representacions acústiques i la seva interacció amb usuaris avaluadors. Aquestes representacions acústiques es basen en models de la percepció humana (Tsuzaki, 2001) que representen millor la resposta auditiva respecte a les mesures de degradació basades en informació espectral de les tècniques d'ajust purament objectives. La distància està limitada dins un interval tancat i fortament discretitzat (els seus valors de similitud són només en el conjunt $\{1,2,3,4,5\}$). No obstant això, aquesta distància és molt restringida, sense complir amb les propietats numèriques de les funcions de distància perceptiva, que són la continuïtat i infinitud (Llorà *et al.*, 2005*b*).

Sistemes perceptius iteratius

Durant els últims anys, la recerca s'ha centrat en nous mètodes capaços d'obtenir conjunts de pesos que mapen els subcostos amb una qualitat auditiva determinada de manera interactiva. Aquests mètodes es poden basar en regressions o altres tècniques més complexes.

La primera aproximació va ser proposada per Wouters i Macon (1998), que van presentar a un grup d'oients conjunts (parelles) de paraules modificant una unitat en cada conjunt. Posteriorment les diferències s'ordenaven segons el seu grau de similitud. El mateix procés s'aplicava, de manera genèrica, per totes les diferències de valors de cada subcost aïlladament. Al final, les dades que s'obtenien reflectien curosament com afectava cada subcost a la qualitat global mitjançant l'esmentada substitució de la unitat.

Downhill Simplex Method Una ampliació de l'esmentada metodologia va ser establerta per Lee *et al.* (2003). En aquest cas es van tornar a sintetitzar paraules monosil·làbiques i a continuació es van ordenar de millor a pitjor amb dues classificacions paral·leles: rànquing de referència (RR) calculat a partir de les preferències de l'oient i rànquing de *test* (TR), que s'obté a partir d'uns pesos heurístics i progressivament actualitzats del sistema. Les classificacions comparen mitjançant una mesura de dissimilitud ad hoc i els pesos s'actualitzen amb el mètode de *Downhill Simplex Method Optimization* (Nelder i Mead, 1965) per tal minimitzar la mesura de dissemblança esmentada.

Així s'obtenen els pesos que guien la selecció d'unitats final. La diferència d'aquest mètode amb els mètodes anteriorment explicats és la capacitat del sistema per predir percepció humana real, i no una mera distància basada en mesures acústiques, o com els mateixos autors l'anomenen, mesures purament físics.

MOS-*Postmapping* A (Chu i Peng, 2001; Peng *et al.*, 2002) reaprofiten els resultats de l'avaluació perceptiva (MOS) de la qualitat final del seu sistema de síntesi per tal de correlar els subcostos de selecció (en el seu cas només lingüístics) amb l'avaluació final dels usuaris. Posteriorment, empren tècniques de regressió tant linials (Chu i Peng, 2001) com d'ordre superior (Peng *et al.*, 2002) per obtenir els pesos més apropiats per tal de millorar la qualitat. Val a dir que en el seu treball varien la mida del corpus de selecció per tal d'obtenir un espectre d'avaluacions prou ampli per a poder fer una millor regressió.

Tomoki Toda, a (Toda *et al.*, 2002, 2006) expandeix la tècnica exposada en el paràgraf anterior avaluant els diferents formes d'integració dels diferents subcostos. Després ana-

litza aïlladament la millor integració dels subcostos mitjançant matrius de correlació parcial. Aquestes matrius de correlació parcial s'obtenen amb proves d'avaluació que adopten intervenció humana. Al final, amb les matrius de correlació parcial s'ajusten els pesos globalment d'una manera semblant als mètodes de regressió lineal.

En aquest sentit, a (Tihelka i Romportl, 2009) s'ajuden de les matrius de similitud parcial per refinar la distància cepstral i que així aquesta simuli millor les preferències humanes a l'hora d'avaluar la bondat del senyal sintètic. Un cop s'obté aquesta distància cepstral acurada, es poden ajustar els pesos de la funció de cost amb els mètodes objectius explicats a 2.4.2. En el seu cas avaluen la correlació entre els diferents subcostos i la matriu de similitud de paraules esmentada, per al final simplement poder descartar els subcostos dolents i quedar-se només amb els bons. De totes maneres, quan avaluen la seva funció de cost global amb una nova matriu de similitud perceptiva, la correlació obtinguda és -0.693 en la seva fita superior, un valor baix de correlació en comparació amb el que obté Toda: -0.84. Els subcostos considerats són diferents mesures de similitud de naturalesa espectral alineats mitjançant *Dynamic Time Warping*.

Una expansió del treball de Chu i Peng (2001); Toda *et al.* (2006) va ser desenvolupada a (Miao *et al.*, 2009) on es combinen algunes de les tècniques que s'han explicat amb anterioritat. L'objectiu del seu treball és trobar un cost de concatenació que combini 8 subcostos definits teòricament i de manera independent, tots ells basats en les trajectòries espectrals (LSF) per concatenar diferents unitats. A tal efecte les diferents distàncies competeixen a través d'un MOS comparatiu (Alvarez i Huckvale, 2002; Sityaev *et al.*, 2006) i converteixen la informació en una avaluació (puntuació) absoluta basada en les votacions (similar al MOS-*Postmapping*). Un cop obtingudes les votacions i normalitzades segons una normalització *z-score* s'aplica una anàlisi de components principals a la matriu $K_{EVAL} \times M_{COST}$ de K_{EVAL} avaluadors i M_{COST} distàncies per obtenir una funció de distància consensuada entre els avaluadors. Al final, per cada al·lòfon es realitza una regressió multilineal entre la matriu PCA de les avaluacions dels usuaris i les distàncies aplicades, per a obtenir un vector de pesos que pondera la importància de cada distància a l'hora de concatenar l'esmentat al·lòfon. Més enllà de les baixes correlacions s'hi afegeix l'inconvenient de la durada de les proves d'ajust, ja que al ser perceptives, a la pràctica, la participació humana en el procés de disseny no té cap garantia d'haver arribat al consens en la seva finalització.

És de rigor observar que la majoria de sistemes de regressió perceptiva no s'han usat per ajustar alhora els pesos de *target* i els de concatenació sinó que el seu ajust ha estat parcial, amb excepció de Toda *et al.* (2006). La majoria dels mètodes explicats, amb excepció de Lee *et al.* (2003), han ponderat diferents subcostos de concatenació tals com la distància

euclídia, la de Mahalanobis (Donovan, 2001), la de Kullback-Leiber (Veldhuis i Klabbers, 2003), la de Itakura-Saito (Rabiner i Juang, 1993), entre d'altres. També s'han pres en consideració diferents paràmetres dins d'aquestes distàncies, com els *cepstrum* (normalment MFCC), (Black i Campbell, 1995; Tsuzaki i Hisashi, 2002; Campillo i Rodríguez-Banga, 2003), els LPC (Macon *et al.*, 1998), o també informació dels formants (Ding i Campbell, 1997), LSF (Vepa *et al.*, 2002) o LSP (Wu i Chen, 2001).

2.5 Línies d'investigació en l'ajust de pesos per sistemes CTP-SU

Resulta evident que l'ajust òptim de la funció de cost és un dels aspectes més importants a tractar en el disseny dels sistemes de CTP-SU. La seva resolució no és trivial, ara per ara, una bona qualitat sintètica s'obté només amb un ajust manual basat en formació o experiència prèvia (coneixement expert).

L'alta qualitat sintètica, per tant, s'assoleix en judicis basats en la prova i error (ajust manual) (Schröder *et al.*, 2009). La principal avantatge de l'ajust manual són les decisions basades en l'experiència: permet prioritzar els millors subcostos només escoltant uns arxius de veu concrets. No obstant això la metodologia és esbiaixada, no es té una visió global de la síntesi i només es basa en una sèrie de percepcions concretes. Aquest biaix es positiu si es pren consciència que alguns errors (selecció d'unitats no òptima) poden ser imperceptibles i en canvi d'altres perceptibles. Tanmateix, segueixen predominant els biaixos cap a l'únic usuari que realitza les proves i les poques frases que s'escolten en l'ajust. L'ajust manual ofereix un bon punt de partida, que permet considerar la sistematització no esbiaixada de l'ajust perceptiu com una línia a seguir.

D'altra banda, la manca d'un model acurat que determini la qualitat (naturalitat + intel·ligibilitat) impedeix aprofitar les tècniques tradicionals d'optimització i cerca que es poden aplicar quan les mesures penalització són objectives (p.ex. simplex descendent). Les aproximacions d'ajust basades en un model de degradació de la qualitat (explícit) estan restringides per la complexitat imposada pel procés de producció de la parla (Hunt i Black, 1996). En altres paraules, l'esmentada tipologia d'ajust sota un model explícit (model matemàtic, p.ex. distàncies cepstrals) provoca la impossibilitat de tenir resultats coherents (correlats) amb la percepció real de la veu sintètica, i per tant, a priori tampoc es pot ajustar automàticament.

Una prova clara de que l'ajust de pesos requereix coneixement expert és la falta d'un criteri únivoc entre si és millor la notació simbòlica (lingüística) o l'acústica sobre els subcostos de *target* de la funció de cost (Taylor, 2006; Campillo i Rodríguez-Banga, 2006; Clark *et al.*, 2007). Tanmateix els errors d'etiquetat de les dues aproximacions denoten incompletitud: la notació lingüística pot resultar ambigua, mentre que la notació acústica és propensa a errors (principalment a causa de l'etiquetat del corpus). L'aparició de treballs que fusionen les dues notacions, per exemple Coorman *et al.* (2000); Campillo (2005); Toda *et al.* (2006); Bonafonte *et al.* (2008), demostra que aquestes parametritzacions no només no competeixen, sinó que també poden cooperar.

Així doncs, la clau per afrontar el problema passa per la definició d'una metodologia capaç de combinar totes les formulacions dels subcostos de forma cooperativa i perceptiva sense la necessitat de coneixement expert. Aquesta metodologia seria de gran utilitat en les fases de disseny o ajust dels CTP-SU per a qualsevol idioma, locutor, estil, etc. per tal d'aconseguir una veu sintètica d'alta qualitat.

L'objectiu principal d'aquesta tesi és trobar vectors de pesos adequats i específics a cada necessitat de la funció de cost. L'ajust hauria de ser mitjançant un procés interactiu sense el coneixement dels paràmetres que intervenen en la selecció d'unitats (metodologia implícita). Els avaluadors no haurien de tenir expertesa sobre els subcostos involucrats en les solucions, i aquestes haurien d'evolucionar en funció de la percepció dels usuaris tractant de trobar de la millor solució per al problema.

CAPÍTOL 3

Computació Evolutiva Interactiva

3.1 Motivacions

Tal com s'ha vist en el capítol anterior (apartat 2.4.3), les aproximacions d'ajust de pesos perceptius mitjançant *simplex* o regressió lineal releguen la participació de l'usuari a una simple validació o bé, en el cas de la regressió, a una participació passiva sense tenir en compte la cooperació activa amb l'algorisme de cerca. La mancança més important però, és que no es té en compte la possible contradicció de l'usuari, bé per manca de criteri, bé per un augment de la fatiga en el transcurs de l'optimització. Per tant, manca un model real que detalli l'evolució de la prova i permeti saber la seva bondat. Això provoca una manca de criteri a l'hora de discriminar quines parelles del binomi prova-usuari afegeixen soroll o no són aptes, i també una falta de criteri sobre com s'han de fusionar els resultats de diferents usuaris.

En aquest capítol s'introdueixen les capacitats de la computació evolutiva interactiva (IEC) per fusionar esforços humans i computacionals amb la percepció subjectiva de l'usuari (Takagi, 2001). En primera instància s'expliquen els fonaments teòrics de la computació evolutiva clàssica i a continuació es detalla la primera aproximació (Alías i Llorà, 2003) evolutiva per ajustar la funció de selecció d'un CTP-SU tot repassant els resultats. Posteriorment s'expliquen els fonaments teòrics de la computació evolutiva interactiva (IEC) i l'adaptació del problema (Alías *et al.*, 2003) de l'ajust de pesos a un escenari d'evolució in-

teractiva, recerca que va iniciar l'estudi de l'IEC dins del grup de recerca. Concretament es van utilitzar algorismes genètics interactius (iGAs) combinant l'ajust de pesos de la funció de cost de selecció amb criteris subjectius. El punt de partida d'aquest enfoc és la incorporació dels criteris subjectius als algorismes genètics clàssics (GAs) mitjançant la substitució de l'avaluació tradicional basada en distàncies computacionals i aplicant un esquema de selecció per torneig (entre dos candidats $s = 2$) (Goldberg, 2002). Per tant la selecció dels millors candidats de la població passa a ser un procés guiat amb intervenció humana.

Aquesta tesi continua el treball previ del grup de recerca enfocat a l'ajust de pesos perceptiu per guiar la selecció d'unitats amb l'ajuda de la computació evolutiva interactiva, que permet per optimitzar la funció de cost d'un CTP-SU sense la necessitat de conèixer els paràmetres involucrats en la selecció d'unitats. Donat que els avaluadors no coneixen el detall de les solucions, en aquest cas els pesos de la funció de cost, simplement evolucionen síntesis perceptivament diferents convergint cap a la millor solució al problema segons el seu criteri subjectiu.

En les primeres aproximacions d'ajust de la funció de selecció basades en iGA (Alías *et al.*, 2003, 2004) es varen detectar diversos problemes: *i*) els usuaris perdien el criteri d'avaluació en el transcurs del procés evolutiu. Al fer-los tornar a avaluar el mateix torneig amb l'ordre canviat, o bé avaluaven un empat o invertien el guanyador; *ii*) no existia cap model, ni tampoc cap indicador, per mesurar la consistència o la validesa del resultat de les proves realitzades; *iii*) no hi havia una metodologia per caracteritzar els resultats d'un usuari si aquest convergia a més d'una solució, ni tampoc cap metodologia per fusionar els diferents resultats més enllà de fer una mitjana simple.

Per superar els esmentats problemes, es proposa una nova metodologia que realitza l'ajust de pesos subjectiu a través d'una tècnica concreta anomenada algoritmes genètics interactius actius (aiGAs -Llorà *et al.* (2005*b*)). Concretament, s'analitza un principi de viabilitat (*proof-of-principle*) on l'evolució esdevé més eficient i fiable que l'estratègia basada en iGAs (Alías, 2006). Aquest fet és degut a l'incorporació d'un model basat en el criteri perceptiu dels usuaris durant el procés d'ajust dels pesos de la funció de cost. La proposta és capaç d'ajustar els diferents pesos (*target* i concatenació) involucrats en el procés de selecció d'unitats, evitant l'heurística que són independents en funció de la seva tipologia (Strom i King, 2008).

3.2 Computació evolutiva i algorismes genètics per l'ajust de pesos

3.2.1 Mètodes de cerca

Els mètodes d'optimització o mètodes de cerca (*search methods*) són populars com a eines de resolució de problemes en àmbits tant diferents com la deducció, la presa de decisions, el raonament de sentit comú, la prova automàtica de teoremes, etc. (Jiménez, 2008). La seva definició és ambigua, no obstant en l'àmbit d'aquesta tesi els mètodes d'optimització numèrica es defineixen com aquells mètodes capaços d'obtenir la millor solució d'una funció quan la seva equació no és coneguda (Goldberg, 1989). Concretament, en el cas de l'ajust de la funció de cost per CTP-SU, la funció desconeguda és la degradació de la naturalitat de la veu en funció d'uns subcostos determinats.

Els models d'optimització numèrica es poden dividir bàsicament en tres grans grups (Goldberg, 1989): *i*) mètodes basats en càlcul, *ii*) mètodes enumeratius i *iii*) mètodes aleatoris. En els propers paràgrafs s'analitzen aquests tres tipus de models.

Els mètodes basats en càlcul poden ser indirectes o directes. Aquests mètodes resolen un sistema d'equacions no lineal que s'obté igualant el gradient de la funció objectiu a zero, perquè en cerquen els extrems locals. En canvi, els mètodes de cerca directes cerquen els òptims locals a través de fer petits salts en la funció i moure's en una direcció determinada pel gradient local (escalat - *Hill-climbing*). Troben la millor solució local i escalen la funció en la direcció del pas més llarg. Malgrat que aquests mètodes s'han estudiat i millorat exhaustivament no es pot garantir la seva robustesa, ja que *i*) els mètodes basats en càlcul són d'abast local; els punts òptims que cerquen són els millors dins d'un veïnatge d'un punt concret i *ii*) els mètodes basats en càlcul depenen únicament de l'existència de funcions derivables (valors de pendent ben definits) en tot el seu domini. Normalment, a la pràctica quan el problema és desconegut, la funció adopta discontinuïtats i espais de cerca multimodals o sorollosos (Jiménez, 2008). Es pot concloure que els mètodes basats en càlcul no són suficientment robustos per funcions desconegudes, o amb problemes de discontinuïtats i problemes de derivabilitat.

Els esquemes enumeratius, malgrat la seva diversitat, parteixen d'una base de funcionament basada en la combinatòria: l'algorisme de cerca prova diversos valors que pren la funció objectiu en els respectius punts dins l'espai de cerca finit, només un en cada instant de temps. Bàsicament, el mètode imita del mètode de cerca combinatori a través de prova i error, que és el més intuïtiu. Malgrat ser algorismes atractius quan el nombre de com-

binacions de les solucions del problema són petites i limitades, quan la complexitat del problema creix aquests mètodes passen a ser molt poc eficients (el WSS explicat a l'apartat 2.4.2 es podria englobar dins d'aquests mètodes).

Finalment, el tercer tipus de mètodes són els mètodes aleatoris, que han esdevingut molt populars un cop reconegudes les mancances dels mètodes anteriors. Tot i així, els passejos aleatoris (*Random Walks*) o altrament anomenats esquemes aleatoris (*Random Schemes*) que cerquen i memoritzen la millor solució, també tenen punts dèbils degut a la seva baixa eficiència. Les cerques aleatòries, en execucions extenses, no obtenen una solució millor que els esquemes enumeratius. De tota manera, s'ha de tenir en compte que no és el mateix parlar de tècniques o mètodes aleatoris que parlar de tècniques o mètodes no deterministes. Un mètode probabilístic (en certa manera aleatori), encara que prengui decisions estocàstiques, en tot moment té una direcció en la cerca, a diferència dels mètodes que exploren l'espai de cerca de manera completament aleatòria. Tal com s'explicarà més endavant, *Simulated Annealing* i la computació evolutiva (EC) són exemples de mètodes amb component probabilística on la direcció de la cerca és guiada. La conclusió on porta tot plegat (Goldberg, 1989) és que tampoc els mètodes de cerca clàssics es poden considerar robustos a priori.

Per acabar la introducció als mètodes d'optimització, és de rigor citar textualment el raonament de Goldberg (1989), que alhora continua el raonament de (Beightler *et al.*, 1979), el qual dóna una primera intuïció sobre quins haurien de ser els objectius a perseguir pels algorismes d'optimització:

"El desig de l'home cap a la perfecció pot trobar-se expressat en la teoria de l'optimització. Aquesta estudia com descriure i assolir el concepte de Millor, un cop se sap mesurar (...) que és Bo i que és Dolent. (...) La teoria de l'optimització engloba l'estudi quantitatiu dels Òptims i els mètodes per trobar-los."

Bàsicament, hi ha dos parts diferenciades a l'hora de definir els objectius de l'optimització numèrica: *i*) l'aproximació eficient a la *ii*) millor solució possible. Normalment es considera la segona part per definir la bondat dels diferents mètodes d'optimització numèrica. Tanmateix, per triar el mètode no s'ha d'obviar l'eficiència de com s'arriba a la solució, ja que a vegades és preferible tenir una bona solució (malgrat no ser la millor) però d'una manera més eficient. Es conclou que l'objectiu més important en l'optimització és simplement la millora contínua (*improvement* o innovació) (Goldberg, 1989). En altres paraules, s'ha d'obtenir un bon rendiment d'una manera computacionalment rellevant en termes de velocitat. Assolir el millor de tots (l'òptim) és poc important en sistemes complexos si per contra s'obté una bona solució de manera ràpida (Goldberg, 2002). En alguns

casos és utòpic buscar la perfecció; tot i així, només es pot aspirar a innovar. És aquí on la computació evolutiva (EC) entra en joc.

3.2.2 Computació evolutiva

La computació evolutiva (EC) basa el seu funcionament en models computacionals inspirats en el caràcter evolutiu que adopta la natura. L'EC engloba un gran ventall de models proposats i estudiats durant les últimes dècades que es coneixen sota el nom d'algorismes evolutius (EA). El seu comportament és no determinista. Això implica que el mateix algorisme pot produir resultats diferents amb les mateixes dades d'entrada ja que les decisions que pren no són deterministes, són aleatòries. Els EA s'inspiren en l'evolució dels éssers vius, adoptant conceptes com la genètica i la selecció natural. En certa manera s'assumeix que la natura sap resoldre problemes d'optimització (origen i supervivència de les espècies) i es busca l'abstracció del seu èxit tant matemàtica com algorítmicament, per així poder aplicar-ho a altres problemes. Des d'un punt de vista d'optimització numèrica, els EA són algorismes que no garanteixen trobar la millor solució al problema però si que asseguren trobar una bona solució (Goldberg, 1989).

L'origen de l'EC es remunta a finals dels anys 50 quan grups de biòlegs van emprar la potència de càlcul que oferia la computació com a eina per a realitzar simulacions de sistemes biològics. Llavors, als anys 60, John Holland es va plantejar la possibilitat d'incorporar mecanismes de selecció natural i supervivència dels millors (*survival-of-the-fittest*) per a resoldre un banc de problemes d'optimització ja resolts anteriorment per la pròpia natura (Jiménez, 2008). La seva recerca tenia dos objectius (Goldberg, 1989): *i*) abstreure i explicar de manera rigorosa el procés adaptatiu dels sistemes naturals i *ii*) dissenyar sistemes artificials en forma de programari que incorporessin les propietats i els mecanismes dels sistemes naturals. Val a dir que aquestes aproximacions han portat a descobriments importants tant per les ciències naturals com per als sistemes artificials (Goldberg, 1989). Tota la seva recerca embrionària es va publicar en el llibre *Adaptation in Natural and Artificial Systems* (Holland, 1975), paradigma de l'EC tal i com la coneixem avui en dia. La computació evolutiva va sorgir bàsicament per una necessitat essencialment acadèmica. El nou paradigma volia concebre un marc més ampli on englobar l'optimització automàtica i així resoldre problemes genèrics que constantment apareixien en el camp de l'optimització. Malauradament, les idees de Holland eren poc eficients a la pràctica (Jiménez, 2008).

A mitjans dels 80, a l'augmentar la potència de càlcul dels ordinadors i aparèixer la informàtica de baix cost i propòsit general (informàtica personal), apareix un nou escena-

ri per a la computació evolutiva. L'EC comença a resoldre amb èxit diversos problemes d'àmbits molt diferents (enginyeria, ciències socials, etc.), difícils de tractar amb anterioritat (Goldberg, 1989; Davis, 1991). L'interès de la computació evolutiva augmenta exponencialment. Aquest fet ve motivat per la capacitat de resoldre problemes reals més enllà de suposicions teòriques (coneixement expert a priori), essent molt motivador el fet que no calgui coneixement expert per resoldre'ls. Amb el temps, els EA passen a ser mètodes imprescindibles que cal tenir en compte respecte d'altres en problemes d'optimització numèrica.

L'èxit dels mètodes d'optimització numèrica es basa en garantir solucions bones, d'una forma robusta i amb un temps de computació assequible. No obstant, també són desitjables altres característiques com la capacitat de resoldre de manera paral·lela problemes de gran dimensionalitat, multimodals, sorollosos o variants en el temps. Per tal de fer front a aquests requeriments resulta essencial que el procediment de cerca es basi en una població de solucions. Per tant, la base teòrica dels EA (evolucionar solucions per assolir una millora global) permet fer-ho d'una manera senzilla. Una cerca basada en població permet intrínsecament el paral·lelisme, que juntament amb els operadors genètics (selecció, herència i mutació) estableix el raonament del Teorema de l'Esquema (Michalewicz, 1992). L'esmentat teorema sosté que les solucions bones explorades per un EA es reprodueixen durant la cerca de manera exponencial. Per tant, el teorema de l'esquema (altrament dit *schema, schemata o Building Blocks*) garanteix el seu bon funcionament (Holland, 1975; Goldberg, 1989).

A diferència d'altres mètodes d'optimització, dissenyar algorismes evolutius no requereix de molta dificultat, ni experiència, ni tampoc coneixement expert. A diferència del mètodes basats en càlcul no necessita molts coneixements matemàtics del problema per al qual està essent dissenyat. Degut al seu caràcter evolutiu, els EA cerquen solucions sense tenir coneixement expert del problema (mètodes febles). Tot i així, els EA adopten una estructura flexible permetent, si es vol, incorporar-hi coneixement expert, cosa que generalment millora l'eficiència del procés quan s'aplica a un problema concret.

No obstant això, els EA presenten l'inconvenient de no ser capaços de trobar la millor solució global degut a diferents fenòmens: epístasis (interferència entre gens especialitzats en diferents aspectes del fenotip), decepció (direcció de la cerca cap a una solució local) o comportament atzarós dels gens (deriva genètica –*genetic drift*) (Goldberg, 1989). Segons el teorema de l'esquema (Michalewicz, 1992) el GA garanteix la propagació exponencial de les millors solucions trobades, però aquest fet pot impedir cercar en totes les solucions possibles i per tant, a vegades, no trobar la millor solució al problema. Gran part de la

recerca per adaptar i millorar els algorismes evolutius s'ha centrat en tractar les diferents solucions d'un problema i establir relacions entre ells (Goldberg i Voessner, 1999).

En resum, els EA es diferencien dels mètodes clàssics en quatre punts (Goldberg, 1989): *i*) els EA treballen amb una codificació (representació) del conjunt de paràmetres, no els paràmetres en sí mateixos, *ii*) Els EA cerquen dins d'una població de punts, no només un punt, *iii*) els EA treballen amb una informació de recompensa (funció objectiva en la majoria de casos), sense emprar derivades ni cap altre coneixement auxiliar, *iv*) els EA empren regles de transició probabilístiques en comptes de regles deterministes.

La EC s'especialitza amb una gran família de tècniques: *i*) Estratègies Evolutives (Rechenberg, 1973), *ii*) Programació Evolutiva (Fogel *et al.*, 1966), *iii*) Programació Genètica (Koza i Poli, 1992) i *iv*) Algorismes Genètics (Goldberg, 1989). Aquests ultims, els algorismes genètics, són els que han copsat més l'atenció general i centrat l'estudi tant teòric com aplicat de l'EC.

3.2.3 Algorismes genètics

Un algorisme genètic (GA) és la representació més clàssica d'un algorisme evolutiu (EA) (Goldberg, 1989). Aquest manté una població de solucions potencials (individus) del problema, les quals evolucionen d'acord als operadors genètics, que són típicament avaluació, selecció, recombinació i mutació. A l'avaluar els individus, de la població cadascun d'ells rep una mesura de bondat *(fitness)* respecte a l'objectiu a aconseguir. La selecció imita el procés de la supervivència en el medi on aquells individus més ben adaptats (*fitness* alt) són els que aconsegueixen sobreviure. La recombinació i mutació modifiquen els individus, proporcionant així varietat i canvi en la població. El funcionament d'un GA es detalla en la figura 3.1. L'algorisme comença inicialitzant i avaluant una població. A cada generació es realitza un procés de selecció dels millors individus (més adaptats, és a dir, amb millor bondat) els quals es constitueixen com a pares de la generació següent, un cop se'ls aplica un procés de recombinació i mutació, donant lloc a un nou conjunt d'individus que a la vegada són avaluats. Es genera una nova generació mitjançant un procés de selecció realitzat entre la població actual $P(t)$ i els nous individus del conjunt, produint una nova població $P(t+1)$. El procés iteratiu acaba quan es compleix una condició de finalització, que normalment s'estableix com un nombre prefixat de generacions. L'algorisme 3.1 mostra l'implementació d'un GA clàssic.

Tal com s'ha explicat a l'apartat 3.2.2, el principal problema de l'implementació clàssica és no poder garantir que es trobi la, o les, millors solucions al problema. Aquest inconveni-

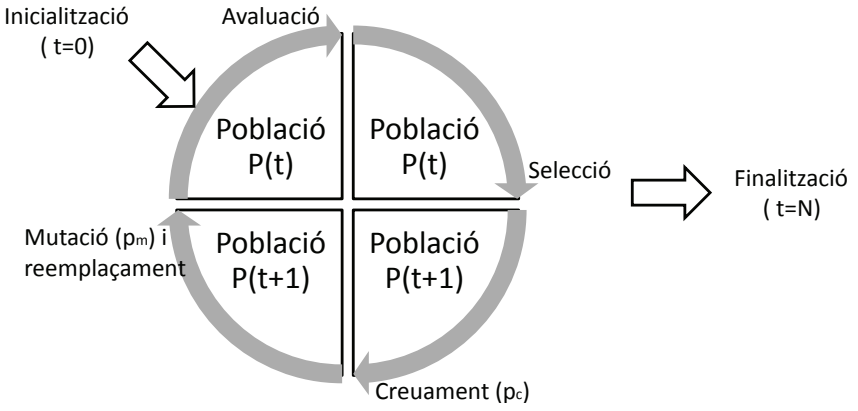


Figura 3.1: Diagrama d'estats d'un algorisme genètic.

Algorisme 3.1 Algorisme genètic.

procedure GA()

1: $t \leftarrow 0$ /* inicialització */
2: Inicialització $P(t)$;
3: Avaluació $P(t)$;
4: **while** ($\neg$ condició de parada) **do**
5: Escollir els millors individus de $P(t)$;
6: $C(t) \Leftarrow$ Recombinació i mutació dels individus de P(t)
7: Avaluació de $C(t)$;
8: Seleccionar els supervivents de $P(t) \cup C(t) \Rightarrow P(t+1)$; /*reemplaçament*/
9: $t \leftarrow t+1$; /* incrementar la generació */
10: **end while**

ent es coneix com el problema de la convergència prematura (Goldberg i Voessner, 1999). De tota manera aquesta dificultat no es limita als algorismes evolutius ja que també apareix en les altres tècniques d'optimització: sovint finalitzen massa ràpidament en un òptim local (Goldberg, 1989).

Quan dins d'una població apareixen individus amb mesures de *fitness* excepcionalment millors que la mitjana de la població, i degut al teorema de l'esquema (Michalewicz, 1992), aquests individus es reprodueixen exponencialment impedint que altres individus contribueixin a nodrir genèticament les subsegüents generacions i com a conseqüència, es perd la diversitat de la informació genètica de la població. Les poblacions passen a ser uniformes i incapaces d'evolucionar (encallades - *stucked*). La recerca realitzada durant les últimes dues dècades ha permès desenvolupar diferents dissenys de GA que impedeixen la convergència prematura. Aquesta recerca es pot classificar en diversos àmbits (Jiménez,

2008): *i*) mecanismes de selecció i mostreig, *ii*) ruptura d'esquemes a causa del creuament, *iii*) fixació de paràmetres i *iv*) característiques pròpies de la funció de *fitness*.

A part de la recerca per evitar la convergència prematura, la recerca dins de l'àmbit dels algorismes evolutius també s'ha centrat en l'adaptació dels GA a diferents tipologies de problemes: restriccions no trivials, funcions multimodals, problemes amb múltiples objectius, etc. (Goldberg, 1989).

Per a fer-ho s'han estudiat diferents mètodes de representació, inicialitzacions diferents de la població, modificacions en les funcions d'avaluació, diferents esquemes de selecció, mostreig i substitució generacional, nous operadors de creuament i mutació, tècniques basades en coneixement específic del problema o en la gestió de restriccions, tècniques per garantir la diversitat, tècniques d'optimització multiobjectiu, etc. (Goldberg, 1989; Michalewicz, 1992). Tanmateix, l'estudi amb detall d'aquestes tècniques queden fora de l'abast d'aquesta dissertació doctoral. Ens els propers apartats simplement es farà un repàs del disseny clàssic de cadascuna de les etapes d'un GA i la seva adaptació al problema de l'ajust de pesos per a la síntesi de la parla basada en la selecció d'unitats.

Codificació: Representació de l'individu

Per aplicar un algorisme genètic a un problema determinat, primerament es codifica el problema en un o diversos cromosomes artificials. Aquests cromosomes artificials poden ser cadenes binàries de 1s i 0s, llistes de paràmetres, o funcions de distribució probabilístiques amb la seva mitjana o desviació típica (Goldberg, 1989), però la clau a tenir en compte és que la maquinària genètica només treballarà amb una representació finita de les solucions, no les solucions pròpiament dites.

Normalment es recomana que tota informació es representi amb uns i zeros segons una codificació binària (Holland, 1975). Per tant, en els algorismes genètics clàssics un individu (cromosoma) és una seqüència de zeros i uns (tira de bits). En el cas de l'ajust de pesos i com a primera aproximació al problema, però, és més adequat treballar amb vectors de nombres reals (discretitzats prèviament). El vector de pesos w tindrà la restricció $\sum w_i = 1$ on w_i son els possibles pesos de la funció de selecció d'unitats (Hunt i Black, 1996).

Per tant, en el problema de l'ajust de pesos es determina que el material genètic (genotip) ve condicionat per un vector de pesos reals els quals condicionen una determinada síntesi o seqüència de veu sintetitzada (fenotip) (Alías *et al.*, 2003).

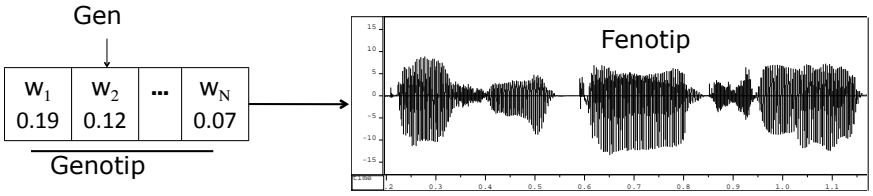


Figura 3.2: Genotip i fenotip d'una possible combinació de pesos.

Inicialització

Un cop escollida la manera de representar els individus, el pas següent és crear la primera generació de la població d'individus. Especificada la mida de la població (el nombre d'individus que tindrà cada generació) es genera un genotip que serà el corresponent a cada individu de la població. En l'algorisme genètic clàssic aquest procés consisteix en generar aleatòriament una cadena de zeros i uns. En el cas dels pesos, es generen valors aleatoris per cada pes (nombres reals).

Avaluació

Abans de poder aplicar el procés de selecció natural es necessita diferenciar els individus més ben adaptats al medi dels pitjor adaptats.

A tal efecte, s'estableix una etapa de gradació de la solucions (avaluació). L'operador d'avaluació assigna a cada individu la probabilitat de sobreviure en el medi. Aquesta probabilitat s'assigna a través de la funció de bondat o, altrament dita, funció de *fitness*. Aquesta pot ser tan senzilla com una funció matemàtica (o un conjunt de funcions) que reprodueix un entorn d'optimització real. Un altre exemple de funció d'avaluació pot ser l'error acumulat (o quadràtic) d'un sistema d'aprenentatge on el procés consisteix en ajustar els seus paràmetres (p.ex. xarxes neuronals, coeficients regressors). Altres sistemes realitzen una avaluació més avançada segons el context de la solució final (p.ex. la distància final recorreguda en un entorn de planificació de rutes).

La idea fonamental és establir una etapa que determini l'aptitud relativa de les diferents solucions al problema, per així poder proporcionar informació a l'algorisme evolutiu de manera que guiï l'evolució de les generacions futures.

Quan aquesta funció d'avaluació resulta impossible de trobar analíticament, o bé presenta una aproximació pobre al problema que tracta de modelar, es pot emprar una funció

d'avaluació de caire interactiu on l'intervenció humana avaluï les poblacions. L'avaluació d'aquests individus es pot realitzar per a cada individu de manera independent (avaluació absoluta), o altrament, inferint una relació de bondat (millor/pijor/igual) relativa respecte altres individus de la població (avaluació relativa). En aquest segon tipus d'avaluació és algorisme qui assigna els valors de *fitness* als individus, a diferència de l'avaluació absoluta on és l'usuari qui assigna directament el valor de *fitness* de l'individu.

Selecció

Un cop s'ha avaluat la població s'escullen els individus més bons per propagar-se a la generació següent. D'aquesta manera s'estableix el procés de *pressió selectiva* inspirat en l'origen de les espècies de Darwin i descrit per Holland (Holland, 1975). Segons aquest procés els individus amb més probabilitat de sobreviure seran els que passaran el seu material genètic a les generacions següents.

En resum, l'operador de selecció condiciona favorablement els millors individus segons la funció de *fitness* com un mecanisme de supervivència del més fort imposat al medi.

El mètodes de selecció més típics són (Goldberg i Deb, 1991): *i*) selecció proporcionada, *ii*) selecció basada en classificació (*ranking*), *iii*) selecció per torneig i *iv*) selecció basada en progenitor (o *steady-state*).

Aquesta dissertació doctoral continua el treball de (Alías i Llorà, 2003) ajustant els pesos d'un CTP-SU mitjançant GA. En el seu treball fan servir selecció per torneig degut a la seva capacitat de treballar amb avaluacions sorolloses de manera eficient (Goldberg, 2002). Cal afegir que la selecció basada en torneig es pot comportar de la mateixa manera que la selecció basada en classificació (*ranking* – (Goldberg i Deb, 1991)). Per aquest motiu s'explica amb detall a continuació:

Selecció basada en *ranking* Baker (1985) va establir el primer disseny de la selecció basada en classificació (figura 3.3). La idea resulta senzilla. S'ordena la població de millor a pitjor, llavors s'assigna la quantitat de còpies de cada individu que s'han de reproduir segons una funció d'assignació, i finalment, es realitza una selecció proporcional d'acord amb aquesta assignació. Grefenstette i Baker (1989) va presentar diverses teories qualitatives relacionades amb aquesta idea, però aquestes teories no assolien el rendiment esperat.

La classificació es realitza en dues etapes. En primer lloc s'ordena una llista d'individus a ser seleccionats, i llavors es consideren els valors d'assignació per a realitzar una selecció proporcional. La computació de la complexitat temporal del procés necessita tractar

Ordre	w_1	w_2	...	w_N	Fitness real	Fitness ordenat
1	0.19	0.12	...	0.07	0.63	1
2	0.14	0.23	...	0.03	0.59	(N-1.5)/N
2	0.13	0.07	...	0.08	0.59	(N-1.5)/N
4	0.09	0.24	...	0.14	0.43	(N-3)/N
⋮	⋮	⋮	⋱	⋮	⋮	⋮
N	0.01	0.10	...	0.07	0.01	1/N

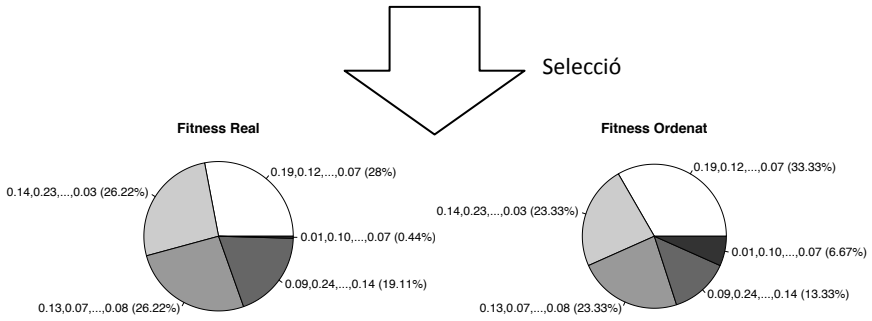


Figura 3.3: Selecció dels individus en un esquema de selecció mitjançant *ranking*.

aquestes dues etapes aïlladament.

L'ordenament de la llista té un cost computacional de $O(n \log n)$, emprant tècniques clàssiques d'ordenació (Goldberg i Deb, 1991). Com a conseqüència, això provoca que la selecció proporcionada trigui entre $O(n)$ i $O(n^2)$. Llavors, s'assumeix que en cap cas el mètode es comporta pitjor que $O(n \log n)$, concloent que la classificació adopta una complexitat temporal de $O(n \log n)$.

Selecció per torneig La selecció per torneig (figura 3.4) es va estudiar per primer cop a (Brindle, 1981), tot i que estudis més recents desenvolupen un sistema de torneig més avançat (Goldberg *et al.*, 1989; Mühlenbein, 1989; Suh i Van Gucht, 1987). La idea tampoc és gaire complexa. S'escullen un nombre d'individus a l'atzar d'una població (amb o sense repetició), llavors es seleccionen els millors individus (guanyadors) de cada torneig per formar part de la generació següent de l'algorisme genètic; aquest procés es pot repetir tantes vegades com es vulgui (típicament fins que el quadre d'aparellaments s'omple). Els tornejos es realitzen normalment amb grups de dos individus (mida torneig $s = 2$), malgrat que també es poden usar tornejos amb un nombre més gran de participants.

El comportament de la selecció per torneig coincideix amb el de la selecció per *ranking*, tant en temps d'execució com en proporcionalitat de les millors solucions (Goldberg i Deb, 1991).

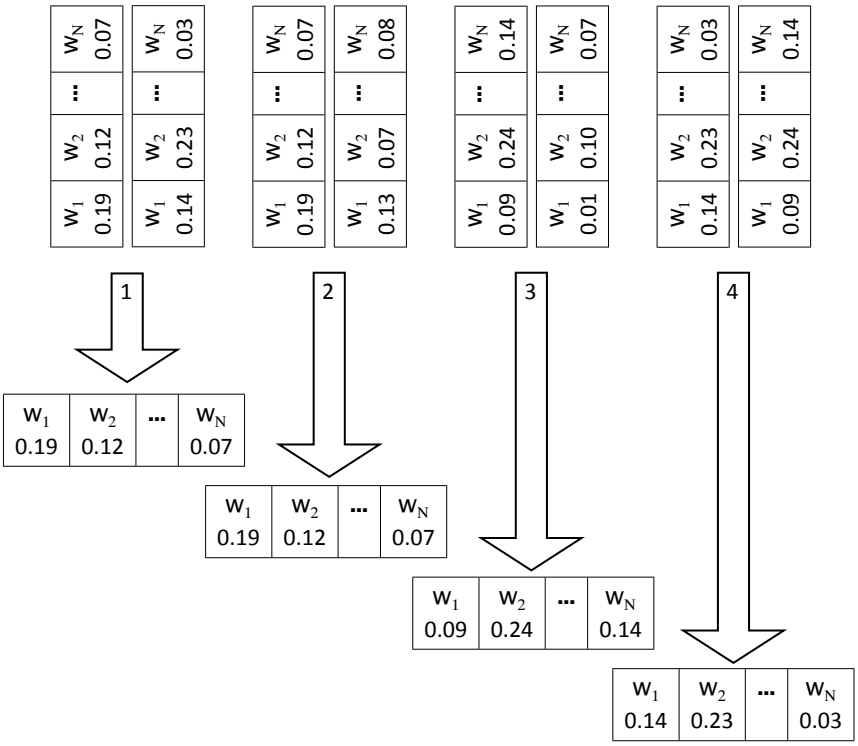


Figura 3.4: Selecció dels individus en un esquema per torneig binari ($s = 2$).

Steady-State La selecció basada en progenitor o *steady-state* va ser introduïda per (Sywerda, 1989) juntament amb el creuament uniforme.

La sel·lecció basada en *steady-state* sempre reemplaça la mateixa proporció (petita) de la població. Els pares no es sel·leccionen de cap manera particular, però sempre considerant la idea que la major part dels cromosomes han de sobreviure a les generacions posteriors (figura 3.5).

El GA funciona de la manera següent: a cada generació *i*) es seleccionen només els millors individus (els pocs que tenen millor *fitness*) per a *ii*) proporcionar el material genètic per les futures generacions (són la base genètica pel creuament i la mutació). Llavors, el nou material creat *iii*) reemplaça els individus (ratio constant) amb pitjor *fitness* de la població. La resta de la població simplement va passant de generació en generació.

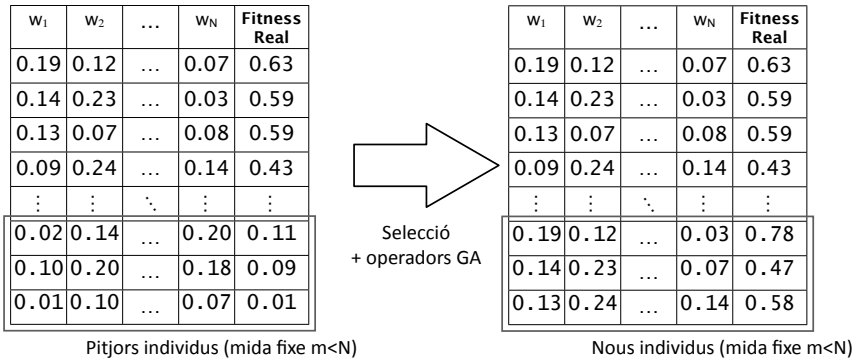


Figura 3.5: Selecció dels individus en un esquema per progenitor (*steady-state*).

Creuament

Només seleccionar els millors no fa evolucionar la població. S'han d'establir mitjans per a poder crear nous individus, i si pot ser millors que els anteriors (innovar). Aquí entren en joc altres operadors genètics. La recombinació és un operador genètic que combina fragments de les solucions existents (pares) per a formar noves solucions (filles) inspirades en el material genètic dels pares i que possiblement seran millors.

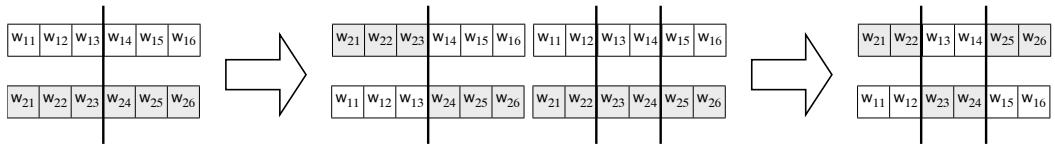
La recombinació (o creuament) normalment escull pares a l'atzar i els creua amb d'altres al voltant d'un punt de creuament totalment aleatori. Cal destacar que es creua tota la població, la idoneïtat d'un individu per a ser recombinat ve donada per un factor probabilístic (Goldberg, 1989), denotat típicament per p_c.

Existeixen diverses tècniques de creuament que es detallen a continuació:

Creuament unipunt: per a cada individu es realitzen els passos següents (exemplificats en la figura 3.6(a)): *i*) S'estableix la idoneïtat de l'individu per ser recombinat segons una probabilitat de creuament. *ii*) S'escull un punt aleatori dins de la tira que representa el genotip. *iii*) S'intercanvien els segments genètics als diferents extrems del punt de tall i s'obtenen nous individus.

Els valors típics de probabilitat de creuament solen estar entre $p_c = 0.6$ i $p_c = 0.7$ (Holland, 1975).

Creuament multipunt: el creuament multipunt segueix bàsicament el mateix procediment que el creuament unipunt. Tanmateix, el creuament multipunt escull dos o més



(a) Creuament unipunt de dos individus.

(b) Creuament multipunt ($p = 2$) de dos individus.

Figura 3.6: Diferents tipus de creuament (unipunt / multipunt).

punts de creuament en els vectors dels pares (pas *ii*) del creuament unipunt). Totes les dades entre els dos punts s'intercanvien anàlogament al creuament unipunt entre els pares, creant dos fills tal i com es mostra en la figura 3.6(b).

Creuament de *cut and slice*: una altra variant de creuament és l'aproximació de "tallar i concatenar" que provoca variacions en la llargada dels vectors fills. Aquesta diferència ve provocada en determinar punts de tall diferents en cadascun dels pares:

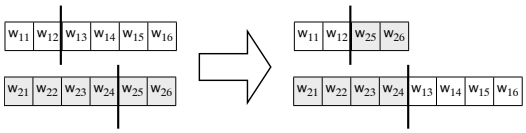
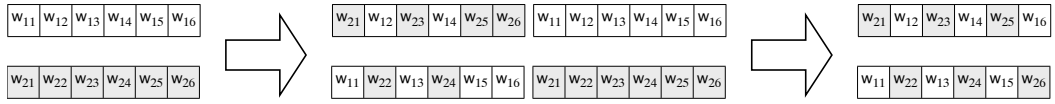


Figura 3.7: Creuament de *cut and slice*.

Creuament uniforme i semi-uniforme: tant en el creuament uniforme com en el creuament semi-uniforme els dos pares es mesclen totalment entre ells per tal de generar els nous fills (no hi ha punt de tall).

En l'esquema de creuament uniforme (*Uniform Crossover - UX*), els bits dels pares s'intercanvien segons una probabilitat determinada (normalment $p_{ux} = 0.5$).

En l'esquema de creuament semi-uniforme (*Half Uniform Crossover* - HUX), es provoca que s'intercanviïn la meitat dels bits diferencials (valors diferents en la mateixa posició per cadascun dels pares). Per a realitzar-ho, es calcula el nombre de bits diferents (distància de Hamming entre els pares) i la meitat exacte d'aquest número són els bits que s'intercanvien en els pares.



(a) Creuament uniforme de dos individus.

(b) Creuament semi-uniforme de dos individus.

Figura 3.8: Diferents tipus de creuament (uniforme/semi-uniforme).

Creuament de cromosomes ordenats: depenent de la representació de dades, és impossible intercanviar les posicions dels gens sense cap més modificació.

Si ens centrem en el problema del viatjant, el *Travel Salesman Problem* (Lin i Kernighan, 1973), el cromosoma es representa amb una llista ordenada de la seqüència de ciutats a visitar. En aplicar els esquemes de creuament explicats anteriorment apareixerien ciutats duplicades o es deixarien de visitar ciutats de la llista. En aquest cas, els cromosomes es mantenen igual fins al punt de tall i posteriorment a aquest, quan s'introdueix el material genètic de l'altre pare, es substitueixen les ciutats duplicades per les ciutats que havien deixat de visitar-se segons l'ordre de l'atre pare. Per exemple, si es vol combinar {0123456789} amb {6543210789} i es limita el punt de creuament en la posició 5, llavors s'obtenen els fills següents: {0123465789},{6543201789}.

Mutació

L'últim operador genètic clàssic és la mutació. Hi ha moltes variants de la mutació, però la idea principal és que els fills siguin idèntics als seus pares, amb excepció d'un o més canvis que pateixen les característiques (codificació genètica) de l'individu. Per sí mateixa, la mutació representa un "passeig aleatori" en el veïnatge de proximitat d'una solució particular. Si la mutació es fa repetidament en una població d'individus, s'esperaria que la població resultant fos indistingible d'una població creada a l'atzar (Goldberg, 1989).

Un cop es disposa de la generació filla, que és fruit en gran part de la recombinació genètica dels seus pares, la mutació s'introdueix variabilitat genètica per així poder inserir nous individus que incorporin trets diferents als seus pares.

El fet de provocar aquesta variabilitat es basa en que , com en la natura mateixa, al recombinar els gens es produeixen certs errors, que són la mutació que pateixen aquests gens per proporcionar noves característiques als individus.

D'altra banda, s'ha de tenir en compte que establir una probabilitat de mutació molt alta impediria per temps la seva selecció i provocaria una mala convergència, o no convergència, de l'algorisme així que se sol treballar amb probabilitats del $p_m = 0.01$ (Goldberg, 1989). Val a dir que 0.01 és un valor orientatiu que varia segons la longitud de la tira de bits.

El criteri que es fa servir per saber si un gen es muta o no és idèntic al criteri fet servir en el creuament. Per a cada gen, es genera un nombre aleatori, si aquest nombre aleatori és inferior a la p_m llavors es muta, en cas contrari no es fa res.

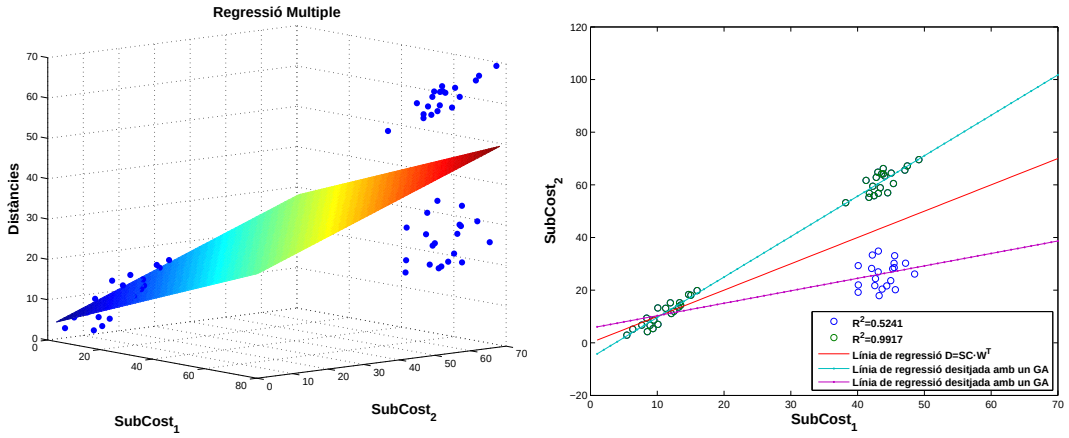
El fet de mutar es basa en introduir una petita variabilitat en la seqüència genètica i això es fa escollint un gen segons el procediment probabilístic descrit anteriorment i aplicant un canvi al seu valor. En el nostre cas el que fem és generar un valor del pes nou totalment aleatori.

3.2.4 Primera aproximació a l'ajust evolutiu de pesos

A (Alías i Llorà, 2003) s'aplica per primer cop la computació evolutiva per a trobar els pesos de la funció de cost de selecció per a la CTP-SU. Concretament, s'apliquen algorismes genètics per obtenir combinacions de pesos (a nivell d'unitat) que optimitzen les distàncies cepstrals de manera no lineal, a diferència de l'aproximació mitjançant regressió multilinear (MLR – veure 2.4.2).

Ajust de pesos amb algorismes genètics per a CTP-SU: motivacions

Tal com s'ha esmentat anteriorment en l'apartat 2.4.2, la tècnica de regressió lineal considera que les dades provenen del mateix origen (en termes de comportament estadístic) i que el seu comportament és lineal. A la vegada considera que la linealitat de les dades és constant per a tot l'espai de cerca (diferents valors de pesos). Això comporta que si existeixen dos conjunts de dades prou diferenciats, la regressió obtindrà un model poc acurat, de compromís, entre els dos conjunts per tal de minimitzar l'error quadràtic. Tal com es pot veure en la figura 3.9(a), una regressió entre dos subcostos i les seves corresponents distàncies cepstrals que presenta aquest comportament provoca que el MLR defineixi un pla que passa entre els dos conjunts de dades. D'altra banda, a la figura 3.9(b) es presenta el resultat d'una regressió entre dos subcostos SC i les seves distàncies d a través dels pesos W^T on es poden veure dues distribucions diferents de dades (en aquest exemple la línia vermella representaria una regressió ideal). Per a cada regressió es dóna el seu paràmetre corresponent R^2 (percentatge de dades explicades pel model – veure apartat 2.4.2). Per la



(a) Exemples de regressions lineals amb conjunts de dades no alineats (3D).

(b) Exemples de regressions lineals amb conjunts de dades no alineats (2D).

Figura 3.9: Exemples de regressions per conjunts de dades no alineats.

que té els conjunts de dades alineats, la regressió té un valor $R^2 \cong 1$, en canvi, l'altra obté un valor de $R^2 \cong 0.5$. És en aquest cas on s'observa la penalització de tenir dos conjunts de dades no alineats (provinents de orígens de dades diferents). Parlant en termes de síntesi de la parla, aquest cas provocaria una funció de cost per a selecció d'unitats ponderada d'una manera poc útil ja que no mapa cap zona de l'espai de cerca, i per tant, empitjora la qualitat de la síntesi respecte una regressió ideal a través d'un dels dos núvols de dades. Altrament, les línies discontínues de punts serien les teòricament desitjables, per una banda, per a tenir una millor mapatge dels subcostos respecte les distàncies cepstrals, i per altra banda, per a poder descartar aquelles unitats que no s'ajusten al model.

Adaptació dels algorismes genètics per al problema d'ajust de pesos per a CTP-SU

L'adaptació de l'algorisme es fa de la manera següent (Alías i Llorà, 2003):

Primerament es realitza una inicialització aleatòria de la població de pesos. Cada individu es representa per un vector $W = (w_1^t, \ldots, w_p^t, w_1^c, \ldots, w_q^c,)$, que és la configuració de pesos, tant de *target* (t) com de concatenació (c), i conté una valors concrets per cada pes.

Cal llavors, avaluar la població. Un individu concret s'avalua segons el seu cost de selecció acumulat per les *k-best* unitats cepstralment més properes, és a dir, usant una distància objectiva. Per tant, per cada configuració de pesos es calcula la mitjana del cost

acumulat al seleccionar les *k-best* unitats (on *k-best*=20 –Alías i Llorà (2003)) més properes en termes de distància cesptral. Per tant, l'equacio 2.4 s'adapta com a funció de *fitness* segons:

$$f(\mathcal{W}) = \frac{1}{k-\text{best}} \sum_{i \in k-\text{best}} C(t_i^n, u_i^n) \tag{3.1}$$

A la generació següent han de sobreviure les millors configuracions de pesos. A l'apartat 3.2.3 s'han repassat els diferents esquemes de selecció. D'entre elles s'utilitza la selecció binària determinista per torneig a causa de la seva habilitat per a tractar amb avaluacions sorolloses d'una manera eficient (Goldberg, 2002), qüestió que cal afrontar al fer un mostreig aleatori dels individus de la població. Un cop s'ha obtingut la nova població, es recombinen els individus en dues fases diferents. En la primera, el creuament, donats dos individus escollits a l'atzar amb una probabilitat p_c, es recombinen els diferents valors de pesos generant així dos nous fills. Aquest procés es fa utilitzant l'operador de creuament per un sol punt (Goldberg, 1989). A més, els nous fills reemplacen als seus pares dins la població. En la segona fase, coneguda com mutació, s'introdueixen variabilitats aleatòries als valors de pesos amb una probabilitat donada p_m (sempre es considera $p_m << p_c << 1$). És llavors quan s'obté una nova població que reemplaça l'original, començant el cicle evolutiu altra vegada. Aquest procés es para quan es satisfan uns certs criteris de parada (en aquest cas, un nombre determinat d'iteracions).

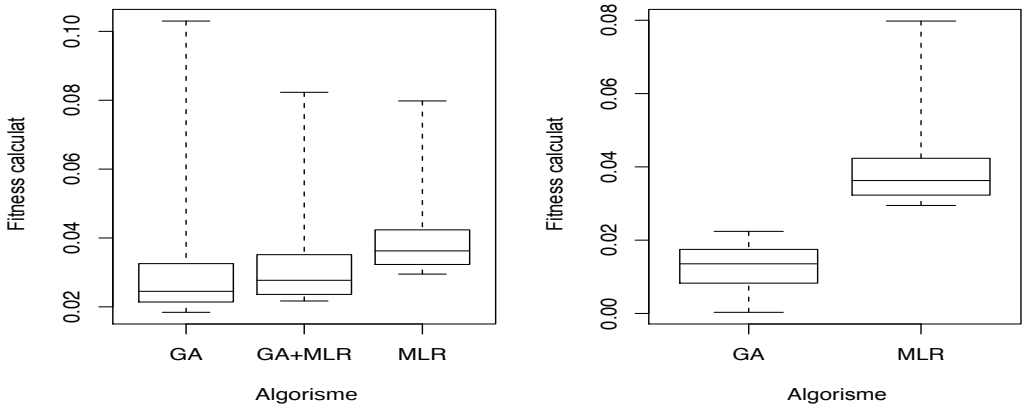
El càlcul del *fitness*, segueix diversos passos. Primerament, s'escull una unitat objectiu aleatòria. Aquest procés de mostreig permet reduir el cost computacional necessari pel càlcul del *fitness* (la funció del cost). En segona instància, es calcula la distància cepstral entre totes les unitats candidates parametritzades i l'objectiu escollit a l'atzar, després d'un alineament temporal mitjançant *Dynamic Time Warping* (DTW) (Ney, 1982). Finalment, les *k-best* millors unitats acústiques (heurísticament assumeixen *k-best* = 20) s'utilitzen per a obtenir el valor final a través de la funció de *fitness*. Aquest valor es calcula com la mitjana de la funció de cost ponderada que implica als *k-best* millors individus recuperats, fent servir els pesos de l'individu w que s'està avaluant (veure l'equació 2.4).

Resultats

El corpus de veu emprat està composat de 1.520 frases en català locutades per un actor professional masculí nadiu amb una tonalitat neutre. La base de dades no és molt extensa (aproximadament 10.000 unitats) per tant no es va realitzar cap poda en el disseny (algo-

risme de *greedy*). Tanmateix, el corpus resulta útil com a marc de proves (*benchmark*) per analitzar un principi de viabilitat (*proof-of-principle*) de la selecció d'unitats mitjançant ajust evolutiu. Les unitats bàsiques del corpus són difonemes i trifonemes (veure apartat 2.3.1). Es de rigor recordar que la qualitat d'un CTP-SU és igual o superior a la d'un sistema CTP per difonemes de segona generació (veure apartat 2.1.2) amb una sola representació per unitat.

Anàlogament a la regressió tipus MLR (Meron i Hirose, 1999), la unitat mínima d'entrenament escollida del sistema són les parelles d'unitats, però en aquest cas difonemes i trifonemes enlloc de frases com a (Meron i Hirose, 1999). Tal com s'ha detallat en l'equació 2.4 s'entrenen alhora els pesos de *target* i els pesos de concatenació. Per tal de veure l'utilitat del GA proposat en entorns d'ajust sorollosos només es van ajustar aquelles unitats amb més de 25 representacions enregistrades en el corpus. Els ajustos evolutius es van realitzar segons el següents paràmetres: mida de la població ($popSize = 200$), número d'iteracions ($iter = 100$), probabilitat de creuament ($p_c = 0.3$), i probabilitat de mutació ($p_m = 0.003$) (Goldberg, 1989, 2002).



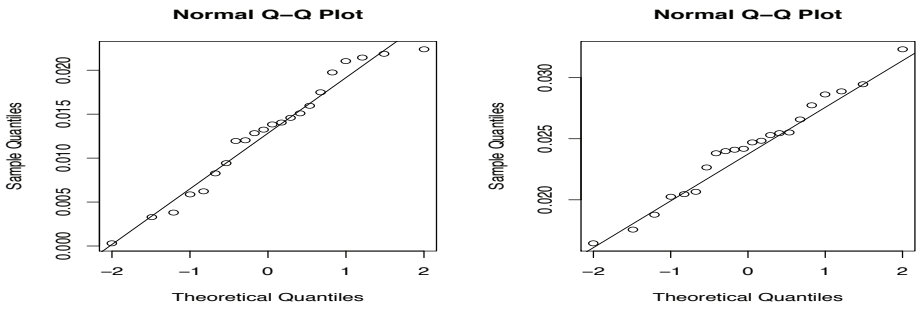
(a) Unitat /b@/.

(b) Totes les unitats.

Figura 3.10: Valors de *fitness* (mínim es millor) obtinguts per totes les unitats segons les diferents metodologies d'ajust automàtic de pesos (MLR-GA) (Alías i Llorà, 2003).

Els autors van escollir la unitat /b@/ (segons notació SAMPA (Wells *et al.*, 1992)) com a unitat-escenari per ajustar el GA i fer una primera comparació dels tres mètodes proposats (MLR, GA i un híbrid de GA amb inicialització del 10% al 50% segons un ajust MLR). La figura 3.10(a) mostra les estadístiques que pren la funció de cost a través de totes les re-

presentacions de la unitat-escenari, considerant la millor configuració de pesos obtinguda per cadascuna de les tècniques. Tal i com es pot observar, el patró de pesos obtingut amb GA obté una millor actuació en termes de cost amitjanat que el MLR, tanmateix amb una desviació més alta. La normalització dels diferents subcosts en l'interval [0,1] es va realitzar segons les funcions d'escalat *max-min* (explicada amb detall més endavant – apartat 4.2.2) degut a que els autors consideren que aquest efecte esbiaixa les solucions obtingudes pel GA. El fet d'inicialitzar del 10% al 50% de la població amb l'execució prèvia d'un MLR només aconseguia minimitzar la desviació dels valors de cost sense minimitzar la seva mitjana, per tant aquesta configuració del sistema d'ajust de pesos fou descartada per fer les proves per a totes les unitats. L'execució repetitiva del GA va obtenir diferents patrons de pesos cosa que confirma el comportament no determinista del GA, en aquest cas degut a una selecció aleatòria de la representació *target* amb la que es calculava la funció de cost. Per tant, al final s'aconseguia un escenari (*landscape*) clarament multimodal degut a l'addició de soroll a la vegada que impedia emprar cap algorisme d'optimització clàssic (apartat 3.2.1). Tanmateix, el GA obtenia bons resultats segons la seva capacitat de ser robust respecte entorns sorollosos donada la impossibilitat d'una execució exhaustiva. Després de la computació dels *fitness* a través de totes les unitats (no només /b@/) (veure figura 3.10(b)) els autors van poder confirmar que les solucions evolucionades pel GA superaven, en termes de mitjana i desviació típica, els valors de la funció de cost resultant obtinguts per MLR.

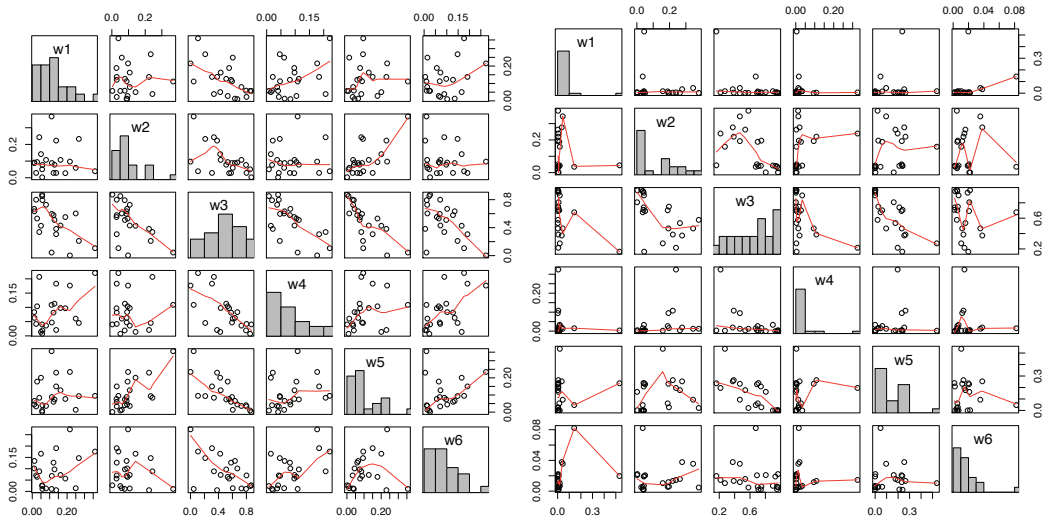


(a) GA. (b) MLR.

Figura 3.11: Comparació quartil-quartil (*qqplot*) dels subcostos assolits a través de totes les unitats pels dos mètodes comparats (Alías i Llorà, 2003).

La funció de cost (veure equacions 2.4 i 3.1) per ambdós algorismes presenten una distribució aproximada a la normal (veure figura 3.11). Per tant, usaren una prova de significància, *t-Student* per veure la validesa estadística dels resultats. La prova de variància

va demostrar que els subcostos de GA eren menors que els subcostos de MLR amb un interval de confiança de $p = 3.756 \cdot 10^{-8}$, dada que confirmava els resultats.



(a) Anàlisi del comportament dels pesos entre ells per a totes les unitats ajustades segons MLR.

(b) Anàlisi del comportament dels pesos entre ells per a totes les unitats ajustades segons GA.

Figura 3.12: Anàlisi del comportament dels pesos entre sí (Alías i Llorà, 2003), on w_1^3 representen els pesos de *target* i w_4^6 els pesos de concatenació.

La figura 3.12 mostra el comportament dels valors dels diferents pesos segons els dos algorismes d'ajust considerats. En el cas del GA només es consideren els pesos al final de l'evolució. La diagonal de les figures mostra l'histograma de cada pes a través de totes les unitats. La resta de subfigures (posicions ij) representen la relació dels valors dels diferents pesos entre si per parelles (w_i,w_j). La línia de suavització sobre-imposada mostra el caràcter d'aquesta correlació: lineal, quadràtica, exponencial, etc. Els resultats mostren que la dependència dels pesos MLR és més lineal que els pesos obtinguts segons GA, tanmateix el seu *fitness* és pitjor (veure figura 3.10). A més, el comportament esbiaixat dels subcostos i el fet de tenir grups d'ajust dependents de la unitat, provoca que $w = 3$ (el cost de *target* de durada) sigui el subcost més rellevant.

En contra, es manifesta que el GA presenta un cost computacional pitjor comparat amb MLR. Tanmateix, el cost computacional creix linealment a diferència de l'aproximació d'ajust per WSS (veure apartat 2.4.2) que creix exponencialment. Es pot concloure que la solució òptima (mínim global) no és impossible de trobar però, el WSS resulta molt costós

computacionalment parlant: per a obtenir bons resultats resulta essencial una discretització intensiva de l'espai de cerca que pot derivar en setmanes o mesos de computació. Això permet que el caràcter elitista del GA pugui assolir bones solucions després de poques iteracions (sense que siguin les millors). Aquesta conclusió confirma la teoria exposada en l'apartat 3.2.1.

3.3 Computació Evolutiva Interactiva per l'ajust de pesos

3.3.1 Algorismes Genètics Interactius (iGAs)

Seguint el raonament fet en la introducció de la computació evolutiva (EC), Hyeduki Takagi (Takagi, 2001) diferencia dos tipus de sistemes a ser optimitzats: *i*) sistemes on el seu rendiment es pot quantificar numèricament - o si més no quantificar - en funcions d'avaluació, i sistemes *ii*) on els seus indicadors d'optimització són difícils d'especificar. Tal com s'ha vist a l'apartat 3.2.1, la gran majoria de la recerca en sistemes d'optimització es basa en l'optimització numèrica la qual maximitza una funció objectiu (o minimitza una funció d'error) i numèricament estableix els diferents passos a seguir en la seva execució: determinació de la direcció de cerca, mida del pas a realitzar, individus que passen a la generació següent, etc.

En el cas dels sistemes on la seva bondat no es pot quantificar numèricament l'única manera de determinar la bondat de les seves possibles solucions és observar com aquestes es comporten en el seu medi final. L'algorisme d'optimització requereix d'una interactivitat amb el medi més enllà de la que li pot proporcionar l'entorn computacional. Normalment, aquesta interactivitat necessita de la intervenció humana de manera directa. En altres paraules, una persona avalua subjectivament la idoneïtat de la solució proposada per resoldre els objectius del problema plantejat.

No obstant això, es fa difícil, per no dir impossible, dissenyar funcions matemàtiques que emulin l'avaluació humana d'una manera explícita. Generalment, les solucions proposades pel sistema són bàsicament de tipus sensorial (imatges, sons, animacions, expressions facials...) i per tant són avaluades des de les impressions, preferències, emocions o nivell de comprensió de l'usuari. Hi ha moltes aplicacions, més enllà de l'òptica artística o estètica, que requereixen d'aquest tipus d'optimització interactiva (Takagi, 2001). Tal com s'ha detallat en el capítol anterior, l'ajust de la funció de selecció per a CTP-SU és un d'aquests casos. Els pesos es poden optimitzar basant-se en l'avaluació subjectiva: l'espai psicològic no es pot aproximar per un gradient, per tant fa falta una altra aproximació que

sigui diferent dels mètodes d'optimització convencionals.

L'adaptació de la computació evolutiva (EC) per a resoldre problemes complexos basats en l'avaluació humana subjectiva porta a la computació evolutiva interactiva (IEC), i en concret, als algorismes genètics interactius (iGAs). Des d'un punt de vista d'implementació, l'algorisme és simplement un GA on la seva funció de *fitness* és substituïda per la intervenció humana. En aquest sentit es pot entendre que l'IEC és un mètode que aporta innovació humana (preferències, intuïcions, emocions, aspectes psicològics) en l'optimització d'un sistema complex (Takagi, 2001). Aquesta cooperació entre EC i l'home es basa en mapar relacions entre les variables a optimitzar i els espais psicològics perceptius. A tal efecte, els usuaris avaluen els individus segons una distància teòrica que s'estableix entre l'ideal (dins la seva noció perceptiva-psicològica) i els candidats reals del sistema. Llavors l'IEC cerca l'òptim global en l'espai de característiques d'acord amb les respostes dels avaluadors.

A diferència de les aproximacions convencionals (explicades a (Nelder i Mead, 1965)) l'IEC no modela les característiques de l'avaluació humana per incrustar-les en un mètode d'optimització existent basat en quantificació numèrica (aproximació analítica). Dins d'una aproximació analítica és molt difícil generar un model de preferències personals. El que fa l'IEC en canvi, és basar-se en una aproximació que "incrusta" l'home com a avaluador substituïnt la funció de *fitness* automàtica (Takagi, 2001) en el mètode d'optimització i permet que els mètodes computacionals optimitzin el sistema mentre que l'home l'avalua.

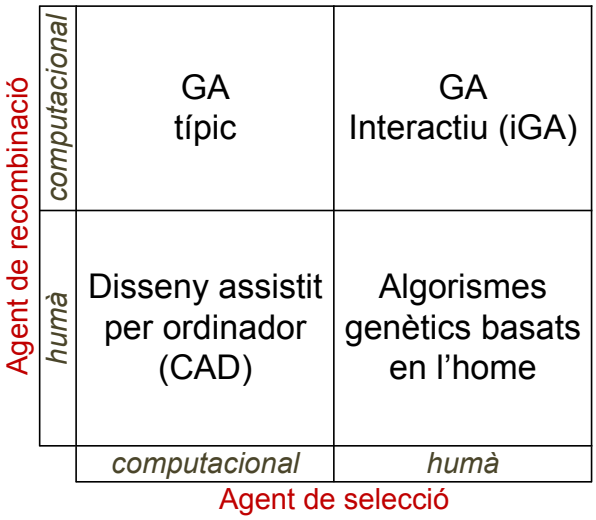


Figura 3.13: Nivells d'interacció home-màquina en termes d'optimització segons (Kosorukoff i Goldberg, 2002).

A Kosorukoff i Goldberg (2002) van fer la classificació de les diferents tipologies de relació home-màquina (*Human-based computation*) a l'hora d'optimitzar sistemes, es detalla a la figura 3.13 i ve determinada per dos eixos: l'eix horitzontal classifica el mètodes segons qui determina la bondat de les solucions candidates (home / màquina), i l'eix vertical classifica els mètodes segons qui realitza la proposta o recombinació de les solucions candidates. Com es pot observar, els iGA són aquells mètodes on el sistema proposa les solucions a avaluar i l'home les avalua.

L'avaluació per part de l'usuari no ve determinada per criteris purament objectius ja que l'avaluació també pot venir influenciada per factors humans aliens a la solució proposada (estat d'ànim, consideracions circumstancials, fatiga, etc.). Les preferències de l'usuari poden canviar en el transcurs de l'execució de les proves igual que pot canviar la seva idea sobre el que és ideal. L'objectiu ideal seria que es pogués determinar la combinació de variables més properes a l'ideal en l'espai psicològic malgrat que l'ideal canvïi en el temps. No obstant això, alguns estudis sobre EC sostenen que la cerca evolutiva és robusta en entorns sorollosos i que les fluctuacions de les preferències de l'usuari influeixen poc en el procés (Ohsaki i Takagi, 1998). Fet posat en dubte per (Llorà *et al.*, 2005*b*).

A diferència dels mètodes d'optimització numèrics, no és fàcil determinar quin és el punt òptim dins l'entorn de l'IEC ja que l'usuari a vegades no pot distingir fenotips lleugerament diferents, mentre que una màquina sí que els podria distingir. En el cas de la síntesi de la parla, tal com s'ha parlat en el capítol anterior, unitats fonètiques diferents poden sonar similars malgrat tenir una parametrització espectral bastant distant: diversos punts en l'espai espectral s'ubiquen al voltant del mateix punt en l'espai perceptiu. Llavors, l'òptim global segons l'IEC no és un punt de l'espai sinó una regió, ja que el propòsit no és determinar un sol conjunt de pesos sinó aquell conjunt de pesos que segueixen un patró el qual és preferit pels usuaris a traves de diferents síntesis.

Existeixen diferents mètodes en els que l'usuari pot interactuar amb la població de solucions dins d'un iGA (Takagi, 2001): *i*) ordenar-les de millor a pitjor, *ii*) mesurar la qualitat absoluta de la solució en una escala de 0 a 100, *iii*) seleccionar un subconjunt de solucions prometedores dins d'un conjunt global de solucions (*pool*). Cadascun d'aquests mètodes combinen diferents mètodes d'avaluació i selecció amb els seus avantatges i inconvenients.

Entre aquestes tècniques cal destacar la selecció per torneig (Goldberg *et al.*, 1989) en ser una de les més àmpliament usades en els esquemes de selecció basats en l'ordre de solucions. Tal com s'ha esmentat a l'apartat 3.2.3, un nombre concret d'individus, s, es selecciona d'una població de mida n. El millor individu d'entre els s individus provoca una còpia en el quadre d'aparellaments. La selecció dels s individus es pot fer amb repetició

o sense repetició. En la selecció sense repetició, l'individu, un cop seleccionat, ja no pot ser candidat a competir en altres aparellaments. En canvi, si la selecció es realitza amb substitució l'individu sí que pot ser candidat a competir en altres aparellaments.

L'esmentat esquema de selecció substitueix la necessitat que l'usuari estableixi una gradació numèrica de les solucions candidates. En canvi, un iGA sota un esquema de torneig amb repetició mostra s solucions candidates, i l'usuari tria la millor solució d'entre les s candidates. La situació més senzilla per a l'usuari és quan $s = 2$. En aquest cas, donades dues solucions s_1 i s_2 en un torneig determinat, l'usuari en tria una, per exemple s_2. Aquesta decisió pot ser adaptada en termes de *fitness* com que 's_2 és millor que s_1', llavors, $f(s_2) > f(s_1)$. Aquesta interpretació introdueix un ordre parcial entre les solucions (Llorà *et al.*, 2005*b*).

3.3.2 Adaptació dels iGAs al problema d'ajust de pesos per sistemes CTP-SU

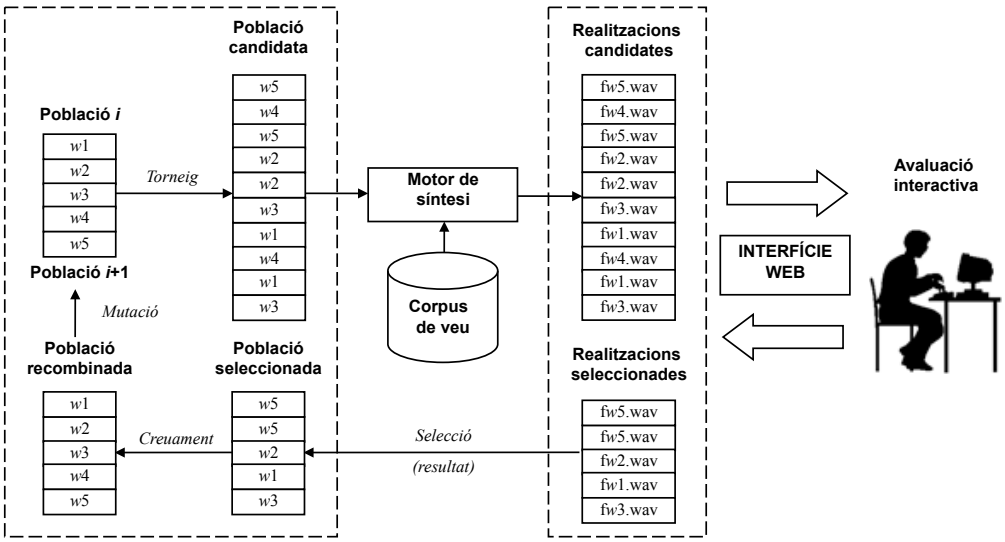


Figura 3.14: Cicle evolutiu de les poblacions de pesos W en l'iGA emprat (Alías *et al.*, 2003).

Els algorismes genètics interactius (iGAs) són un model d'optimització capaç de combinar l'ajust de paràmetres quantitatius amb l'avaluació subjectiva dels resultats. Aquests tipus d'algorismes han estat aplicats en gràfics per ordinador, en enginyeria mecànica, o en processat del senyal, entre d'altres (Takagi, 2001), així com en sistemes de processament de la parla (Watanabe i Takagi, 1995; Sato, 1997; Todoroki i Takagi, 2000). En alguns d'ells, els iGAs han estat utilitzats per l'ajust de coeficients en filtres FIR. En altres treballs (Alm i Llora, 2006) els iGAs han estat emprats per l'ajust de paràmetres de control per a la

incorporació d'emocions al senyal de veu.

A (Alías *et al.*, 2003) es dissenya un iGA per realitzar el mateix ajust de pesos dels subcostos de la funció de selecció descrit en l'apartat 3.2.4.

L'estructura de l'algorisme segueix la mateixa estructura bàsica que es pot trobar en un GA convencional (Goldberg, 1989, 2002). Es basa en l'evolució d'un vector d'individus $W = (w_0, \ldots, w_n)$ que es correspon als pesos emprats en la funció de cost, fins a trobar la seva configuració $W = (w_1^t, \ldots, w_p^t, w_1^c, \ldots, w_q^c)$ òptima.

El procés d'evolució es divideix en dues fases: *i*) la selecció de les millors solucions contingudes en la població, i *ii*) la seva posterior recombinació per generar noves solucions (vegeu la figura 3.14). És en el procés de selecció on els iGAs presenten les seves peculiaritats. En cadascuna de les iteracions del procés d'ajust, l'algorisme disposa d'un conjunt de pesos w_i per a la síntesi del text estudiat (CTP-SU). El resultat d'aquesta síntesi és avaluat interactivament per l'usuari al que se li presenten les possibles solucions de la iteració segons un procés de selecció per torneig binari. D'aquesta manera, l'usuari escollirà, després d'escoltar les vegades que necessiti les dues solucions de la parella, aquella opció (vector de pesos w_i) que condueixi al millor senyal de veu des d'un punt de vista subjectiu.

Pel que fa al procés de recombinació de l'iGA, aquest segueix la mateixa directiva que en un algorisme genètic tradicional. Aquest procés es basa en la recombinació probabilística del material genètic (en aquest cas el conjunt de pesos W). El mecanisme utilitzat consisteix en l'intercanvi de fragments de material genètic procedents de dos progenitors (població seleccionada). En aquest cas, l'operador d'encreuament utilitzat és el clàssic de punt d'encreuament únic (Goldberg, 1989). Aquest procés ve acompanyat per la incorporació de possibles errors en aquest procés de recombinació (mutació). Això s'aconsegueix pertorbant probabilísticament fragments del material genètic, afectant alguns dels pesos (Goldberg, 1989).

L'implementació de l'esmentat iGA per a l'ajust de pesos es va dur a terme a través d'una plataforma web basada en tecnologia LAMP (*Linux, Apache, MySQL i PHP*). Els detalls tècnics d'aquesta plataforma es poden trobar a Formiga (2003).

3.3.3 Experiments i resultats

Els experiments es van realitzar sobre el mateix corpus en català detallat en l'apartat 3.2.4. El seu objectiu era: *i*) estudiar el principi de viabilitat (*proof-of-principle*) tant de la plataforma desenvolupada com de l'ajust subjectiu de pesos mitjançant iGAs, i *ii*) obtenir les

condicions que caracteritzen l'escenari d'un ajust interactiu dels pesos que guien la funció de selecció d'unitats en un entorn CTP-SU.

L'algorisme evolutiu es va dissenyar amb una mida de la població de 15 individus, on la seves probabilitats de creuament i mutació eren $p_c = 0.6$ i $p_m = 0.01$. Els iGA segueixen el mateix funcionament que un GA amb selecció per torneig i creuament uniforme (Takagi, 2001), per tant, segons l'equació de diferències per calcular el temps de convergència explicada a Goldberg i Deb (1991), el temps de convergència per una població de 15 individus és $\log(15) + \log(\ln(15)) = 5.3341$, fent necessari un mínim de 5 generacions per obtenir resultats fiables a nivell teòric. Tres usuaris experts van avaluar 5 frases extretes d'un documental de televisió. L'informació d'entrada del sistema de síntesi consistia en una transcripció fonètica, la seva prosòdia associada (en termes de *pitch*, energia i durada) i uns valors concrets dels pesos de la funció de selecció d'unitats. Val a dir que es feia servir la prosòdia original en el corpus per tal d'optimitzar la selecció d'unitats de manera insensible a la propagació d'errors esmentada en l'apartat 2.3.4. A cada pas del procés iteratiu (d'un total de 7 iteracions), l'usuari escollia el millor individu de dos possibles candidats a partir dels resultats sintètics que produia. Cal esmentar que a l'usuari se li oferia l'opció d'escoltar la frase original per tenir una referència respecte els resultats sintètics.

Els resultats obtinguts van permetre extreure una sèrie de conclusions que van conformar l'escenari de punt de partida per l'ajust perceptiu dels pesos de la funció de cost mitjançant IEC.

En relació al procés de realització de les proves, segons enquestes informals es van obtenir les consideracions següents:

i) El criteri d'optimització dels usuaris canviava durant el transcurs de les proves: malgrat comencessin l'ajust segons un determinat criteri, durant el transcurs de les proves el criteri canviava inconscientment i l'usuari es sorprenia a ell mateix contradient-se.

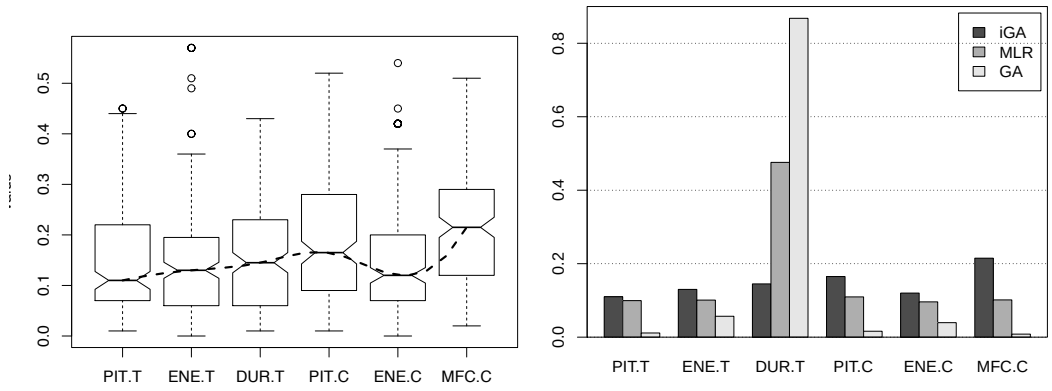
ii) Usuaris diferents adoptaven criteris perceptius que, en alguns casos, podien resultar contradictoris en funció de la seva expertesa (p.ex. l'entonació prosòdica era més o menys important que la qualitat de les unions segons l'usuari).

iii) En determinades frases i després d'un cert nombre d'iteracions, la diferència entre els candidats (frases sintetitzades) era pràcticament imperceptible, de manera que la prova esdevenia tediosa sense aportar nova informació (havia convergit i per tant resultava estancada).

iv) Un corpus com el que es va emprar, de reduïda dimensionalitat ($\approx$ 15 mins de veu), re-

sultava l'escenari més complicat per l'ajust de la funció de selecció degut a les poques realitzacions per unitat disponibles.

v) El fet que les frases de referència (originals) pertanyessin a un locutor diferent respecte la veu sintètica dificultava la tasca de l'usuari en l'avaluació de les solucions. Així doncs, es va concloure que és convenient mantenir la homogeneïtat en les locucions per poder dur a terme una comparació més precisa considerant aspectes implícits al parlant com el ritme i l'entonació de la parla.



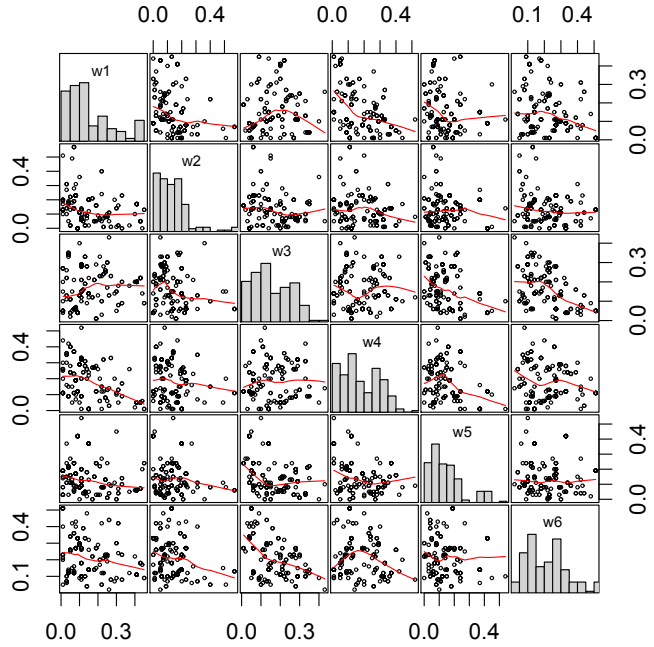
(a) *Boxplots* dels valors dels pesos al finalitzar les 7 generacions de les proves.

(b) Comparació dels pesos obtinguts en aplicar iGA amb els altres mètodes automàtics.

Figura 3.15: Resultats obtinguts amb iGA durant 7 generacions.

Si es compara la tendència dels valors dels pesos respecte els mètodes purament automàtics (figura 3.15(b)) es pot veure que les diferències de l'iGA són més suavitzades que respecte la no uniformitat dels altres mètodes (MLR, GA). El pes de durada DUR.T passa a tenir un valor baix en comparació amb els altres mètodes. En canvi, els valors dels pesos de concatenació, bàsicament PIT.C i MFC.C, adopten la major importància en termes generals. En aquest sentit es pot considerar que la tècnica*Dynamic Time Warping* emprada en les tècniques automàtiques no aconsegueix relativitzar l'error respecte la diferència de durades ja que propaga l'error espectral per totes les trames de la diferència. A part, es pot observar que el mètode de comparació mitjançant distàncies cepstrals no és sensible a les discontinuïtats de concatenació (artefactes) que són prioritzades per part de l'oient. Per últim, l'eficiència de les distàncies cepstrals per a poder realitzar l'ajust de pesos no pot ser validada ja que en l'avaluació interactiva els valors assolits són diferents.

A la figura 3.15(a) es pot observar el patró que segueixen els diferents valors de pesos



(a) Gràfic del comportament dels diferents subcostos w entre si en el transcurs d'un iGA.

$$R_w = \begin{pmatrix} 1 & -0.24 & 0.16 & \mathbf{-0.45} & -0.22 & -0.27 \\ -0.24 & 1 & -0.22 & -0.15 & -0.14 & -0.21 \\ 0.16 & -0.22 & 1 & 0.04 & \textit{-0.35} & \mathbf{-0.50} \\ \mathbf{-0.45} & -0.15 & 0.04 & 1 & \textit{-0.31} & -0.14 \\ -0.22 & -0.14 & \textit{-0.35} & \textit{-0.31} & 1 & 0.0092 \\ -0.27 & -0.21 & \mathbf{-0.50} & -0.14 & 0.0092 & 1 \end{pmatrix}$$

Figura 3.16: Correlacions lineals entre els valors dels pesos després de finalitzar les 7 generacions de l'iGA, on w_1 =PMG.T, w_2 =ENE.T, w_3 =DUR.T, w_4 =PMG.C, w_5 =ENE.C, w_6 =MFC.C.

en l'última generació considerant les 5 proves i els 3 usuaris. Es pot veure, en termes generals, que les diferències no resulten significatives amb les excepcions de l'importància que adopten els *Mel-Frequency cepstrum* en concatenació (w_6 =MFC.C) i el *pitch* de concatenació (w_4 =PMG.C) que indica la importància que té la continuïtat de la f_0 entre les unitats.

Si s'analitza el comportament dels pesos entre sí (figura 3.16(a)) es pot observar que, en termes generals, no hi ha cap evidència de comportament lineal entre les unitats tal i com passava en el treball detallat en l'apartat 3.2.4 (Alías i Llorà, 2003). El que sí que és de rigor destacar és la presència d'unes lleugeres correlacions negatives en la comparació per parelles entre (w_3 =DUR.T, w_6 =MFC.C)=-0.50 i (w_1 =PMG.T, w_4 =PMG.C)=-0.45, les quals indiquen la competència d'aquests paràmetres per sobreviure pels mateixos recursos. En efecte, si es dóna un cop d'ull als histogrames (diagonal de la figura 3.16(a)), es pot observar que els histogrames de w_3 =DUR.T, w_4 =PMG.C i w_6 =MFC.C adopten certa bimodalitat que impedeix adoptar una sola solució com la millor, ja que aquesta està distribuïda en dos màxims locals. La tria d'un o altre màxim local pot venir condicionada per diversos motius (p.ex. la tipologia de les unitats a sintetitzar) però la primera conclusió és que les solucions obtingudes no es poden mitjanar ja que el valor resultant pot caure en un mínim en comptes d'un màxim local. L'altre conclusió a observar és que les esmentades correlacions negatives porten a pensar amb l'idea d'un front de solucions òptimes en comptes d'una solució única (enfocament multiobjectiu). En altres paraules, es pot concloure que hi ha certs aspectes de multimodalitat i multiobjectivitat que cal afrontar per tal d'avançar en la recerca en aquest camp.

3.4 Combatre la fatiga i la robustesa dels iGAs: iGAs actius (aiGAs)

En aquesta apartat es presenta una evolució dels iGAs proposada per (Llorà *et al.*, 2005*b*) i anomenada algorismes genètics interactius actius (*active interactive Genetic Algorithms* o també anomenats aiGAs) amb l'objectiu de millorar l'eficiència de l'intervenció de l'usuari en els iGAs. Aquesta nova evolució permet reduir la fatiga en el transcurs del procés interactiu i alhora millorar la fiabilitat de les solucions finals.

3.4.1 Obtenció activa vs. obtenció passiva de les preferències de l'usuari

Aprenentatge actiu

Per entendre el concepte actiu, primerament s'ha d'entendre quin rol adopta aquest concepte dins de l'àmbit de l'aprenentatge artificial. Un dels primers treballs que sintetitza els diferents treballs basats en aprenentatge actiu en models estadístics (dels quals l'aprenentatge artificial en forma part) és el treball de Cohn *et al.* (1996) el qual diferencia el terme d'aprenentatge clàssic o *passiu* d'aprenentatge *actiu*, en el qual els exemples d'aprenentatge són costosos i de difícil obtenció.

Segons la seva diferenciació, hi ha dos tipus de rols en el model d'aprenentatge estadístic: *i*) la gran majoria de sistemes d'aprenentatge artificial tracten l'elaboració del model com un agent passiu (*passive recipient*) respecte les dades que s'han de processar. Aquesta aproximació "passiva" ignora el fet que, en moltes situacions, l'eina més potent del model d'aprenentatge és la seva capacitat d'actuar, recopilar dades, i per tant, d'influenciar respecte les preguntes sobre l'entorn que intenta entendre. Això dóna lloc a una nova generació de *ii*) sistemes d'aprenentatge "actiu" que estudien com treure profit d'aquesta habilitat en els diferents mètodes d'aprenentatge.

D'una manera formal, el mètodes basats en l'aprenentatge actiu estudien el fenomen del bucle tancat (*closed-loop*) que es dóna quan un model d'aprenentatge selecciona o realitza preguntes que influencien quines dades conformen el conjunt d'entrenament. Un exemple típic d'aprenentatge actiu és l'endevinalla de les 20 preguntes (Taylor i Faust, 1952).

Si un model d'aprenentatge és capaç de fer les preguntes correctes, les restriccions sobre el grup de dades d'entrenament poden reduir-se d'una manera dràstica i, fins i tot, alguns problemes d'aprenentatge que resulten amb una complexitat temporal *NP-Hard*

poden esdevenir temporalment polinomials (Angluin, 1988; Baum, 1991). A la pràctica, l'aprenentatge actiu ofereix les millors prestacions en aquelles situacions on l'obtenció de dades és complex, lent o sorollós (Cohn *et al.*, 1996).

Seguint el recull fet per (Cohn *et al.*, 1996) es pot observar que s'han establert diverses heurístiques per triar quins punts conformen el conjunt d'aprenentatge en diversos escenaris quan: *i*) no es disposa de conjunt d'entrenament (Whitehead i Ballard, 1991), *ii*) el conjunt d'entrenament es dissenya per provar errors en el classificador (Linden i Weber, 1993), o bé les dades d'entrenament són poc fiables (Thrun i Möller, 1993), *iii*) el model és variant en el temps (Cohn *et al.*, 1994), o per últim, *iv*) es disposa de dades a priori que generen aprenentatge previ (Schmidhuber i Storck, 1993).

Bàsicament, els mètodes d'aprenentatge actiu seleccionen les dades d'entrada a processar per tal de minimitzar l'esperança de la variança de l'error del model d'aprenentatge segons l'equació 3.2 (Cohn *et al.*, 1996).

$$\left\langle \widetilde{\sigma}^2_{\widehat{y}} \right\rangle = E_{\mathcal{D}\cup(\widetilde{x},\widetilde{y})} \left[\sigma^2_{\widehat{y}} | \widetilde{x} \right] \tag{3.2}$$

on $\langle\cdot\rangle$ és l'equivalent de l'esperança $E_{\mathcal{D}}[\cdot]$ donat un punt fix x en el conjunt d'entrenament $\mathcal{D}$. Quan es selecciona un nou punt d'entrenament $\widetilde{x}$, es computa el seu valor resultant $\widetilde{y}$ i la parella de valors $(\widetilde{x},\widetilde{y})$ s'afegeix al conjunt d'entrenament provocant la variació de l'error de predicció del model d'entrenament $\sigma^2_{\widehat{y}}$.

Mètodes de cerca activa

Dins l'entorn de l'optimització numèrica (apartat 3.2.1) es poden fer servir les avantatges de l'aprenentatge actiu per guiar la cerca de qualsevol mètode d'optimització numèrica. En concret, dins l'entorn d'optimització, el model d'aprenentatge seria qui guiaria la cerca dels punts a explorar ja que l'algorisme realitzaria la predicció de la probabilitat dels possibles punts per tal de maximitzar o minimitzar la funció del problema. Un dels primers exemples d'optimització guiada per aprenentatge artificial es pot trobar en la metodologia de resposta en superfície (Box i Draper, 1987), en la qual un algorisme d'aprenentatge artificial guia el procés de *hill-climbing* a través de l'espai d'entrada.

El sistema anàleg en computació evolutiva interactiva (IEC) seria un procés on un mecanisme d'aprenentatge artificial guiés quins punts han de ser avaluats per l'usuari. Aquests punts s'escollirien de manera intel·ligent per tal d'augmentar l'eficiència de la interacció amb l'usuari i per tal de reduir la seva fatiga.

En els propers apartats es detalla el recorregut explorat per Llorà *et al.* (2005*b*) per tal d'obtenir un iGA amb comportament actiu (*active interactive Genetic Algorithm* - aiGA) per així poder obtenir un ajust de pesos en CTP-SU de manera més eficient.

3.4.2 Problemes d'esdevenir interactiu: fatiga i robustesa

El problema principal dels iGAs és la fatiga humana, comuna a tots els sistemes d'optimització que inclouen interacció home-màquina en el procés d'ajust interactiu. La primera intuïció per a solventar aquest problema passa, o bé per accelerar la convergència de l'EC amb una mida de població més petita, o bé per disminuir el nombre de generacions. De totes maneres, aquestes optimitzacions comporten una pitjor qualitat en les solucions trobades durant la cerca (Llorà *et al.*, 2005*b*). Des d'un punt de vista de disseny, el nombre d'estímuls en una avaluació usant un iGA ve limitada per la capacitat humana de recordar diferents solucions simultàniament, en el pitjor dels casos s'assumeix el cas més restrictiu possible $s = 2$. A menys capacitat d'avaluar diferents solucions a la vegada, augmenta el nombre necessari de generacions evolutives. Tanmateix, el nombre de generacions dels iGAs ve limitada per la fatiga humana: típicament una cerca interactiva amb 10 o 20 generacions per part de l'usuari és el nombre màxim de generacions limita la cerca mitjançant un IEC (Takagi, 2001).

Per tal de resoldre la problemàtica de l'IEC comparada amb l'EC clàssica (no interactiva), a Takagi (2001) es va presentar una primera revisió de les línies de recerca que s'havien de seguir en els iGAs. Aquestes línies, vigents avui en dia, de recerca incloïen: *i*) mètodes d'interacció que permetin discretitzar el *fitness* (l'interacció amb l'usuari no és contínua ja que és finita i acotada), *ii*) la predicció dels valors de *fitness*, *iii*) millorar l'interfície per a tasques dinàmiques, *iv*) l'acceleració de la convergència dels iGAs, *v*) combinació de la computació basada en computació evolutiva i amb altres tipus de computació (p.ex. aprenentatge artificial), *v*) metodologia de cerca activa, i *vi*) establir una teoria en la cerca interactiva.

A (Llorà *et al.*, 2005*b*) es reorganitzen les línies de recerca dels iGAs proposades per Takagi a cinc elements principals per tal d'obtenir solucions robustes d'una manera efectiva:

i) *Definició clara de l'objectiu:* una descripció precisa de l'objectiu és clau per ajudar a l'usuari a obtenir un procés d'innovació satisfactori. Una definició clara ajuda a evolucionar cap a solucions d'alta qualitat. A més, si aquesta definició es manté a través de l'execució, l'ambigüitat de l'usuari se simplifica molt.

ii) *Impacte del problema de la visualització:* convé que les solucions presentades a l'usuari siguin senzilles, comprensibles i comparables. Si la presentació de solucions a l'usuari és massa ambigua l'usuari es perdrà amb els detalls. Si no hi ha una manera simple de comparar les solucions qualitativament, l'usuari no serà capaç de prendre les decisions adequades. Si aquesta comparació qualitativa no és senzilla, la qualitat de les solucions de l'usuari decreixeran ni penalitzaran, i molt, l'actuació dels iGAs.

iii) *Manca d'un fitness real:* els iGAs no disposen d'una funció de *fitness* quantitativa anàloga a les usades pels Algorismes Genètics (GAs). La natura qualitativa del procés d'avaluació normalment porta a escenaris on es demana a l'usuari que puntuï possibles solucions de cara a uns subconjunts de solucions seleccionades.

iv) *Fatiga:* la fatiga de l'usuari és un element crític per produir solucions d'alta qualitat de l'algorisme. Quant més llarg sigui l'interval de temps per assolir la convergència més problemes de concentració tindrà l'usuari. La fatiga passa a ser la raó principal d'una parada precipitada dels iGAs i, malauradament, comporta solucions de baixa qualitat.

v) *Persistència del criteri de l'usuari:* l'usuari pot canviar el seu criteri d'avaluació a través de l'algorisme genètic interactiu (iGA) provocant una convergència sorollosa. El criteri de l'usuari pot patir canvis al llarg de l'execució i això provoca treballar en escenaris d'optimització dinàmics. Un element clau pels iGAs és l'establiment de mecanismes que ajudin a l'usuari a mantenir el seu criteri a través del procés iteratiu d'interacció.

En la seva proposta Llorà *et al.* (2005*b*), proposen una nova aproximació al problema de l'ajust interactiu usant GAs basada en proporcionar una funció de *fitness* real per guiar la cerca i alhora reaprofitar les avaluacions comparatives proporcionades per l'usuari per establir un model que representi l'espai psicològic de l'usuari. Aquesta nova proposta s'anomena iGA actiu o aiGA.

3.4.3 Funció de *fitness* sintètica i grafs d'ordre parcial

Per tal de superar la fatiga de l'usuari s'han abordat diverses tècniques que intenten millorar l'eficiència dels GAs (Goldberg, 2002). A (Goldberg, 2002; Sastry *et al.*, 2004) es presenten diverses tècniques (paral·lelització, continuació temporal, relaxació de l'avaluació i hibridació) que permeten reduir el nombre d'avaluacions reduint alhora el temps de convergència del GA.

No obstant això, les millores centrades en l'eficiència assumeixen l'existència d'una funció de *fitness* objectiva (sense intervenció humana) i per tant, que el seu càlcul no pro-

dueix cap tipus de fatiga en el transcurs del temps. Aquesta assumpció pot no complir-se en el paradigma de l'IEC.

D'altra banda, la naturalesa pròpia de l'IEC no permet quantificar, de manera robusta (*fitness* global) les avaluacions dels iGAs (Llorà *et al.*, 2005*b*). Això genera una sèrie de qüestions que necessiten de resposta. Per exemple, si una execució de l'iGA es fa mitjançant dos avaluadors paral·lels amb diferent criteri, llavors l'iGA ha de ser capaç de combinar els diferents criteris d'avaluació subjectiva. A més, no es pot assumir de manera teòrica la coherència del criteri d'avaluació a través d'avaluadors paral·lels: diferents usuaris poden portar a diferents solucions. Llavors, per abordar aquests problemes, es poden recuperar tècniques clàssiques d'EC per a problemes multimodals o multiobjectius.

En tot cas, malgrat la millora de l'eficiència, per reduir la fatiga i, per tant, per millorar la qualitat de les solucions obtingudes, es fa necessari obtenir una funció de *fitness* automàtica inspirada en l'avaluació real. Seguidament s'exposen les característiques que hauria de seguir l'esmentat *fitness*.

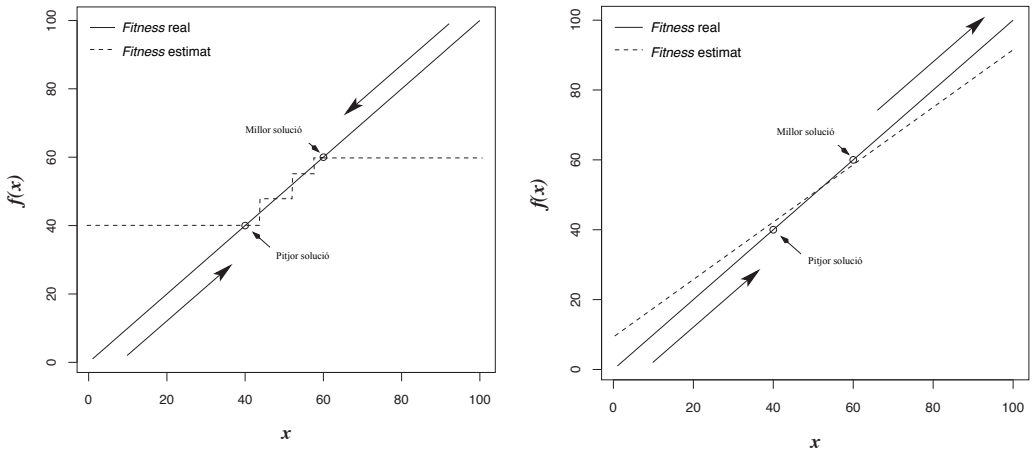
Prerequisits del *fitness* sintètic

Qualsevol intent de generar un *fitness* subjectiu s'hauria de basar en l'ordre parcial que l'usuari infereix a les solucions candidates. A més, la funció de *fitness* subjectiu hauria de complir, almenys, dues propietats (Llorà *et al.*, 2005*b*): *i*) extrapolació dels seus valors més enllà dels de l'entrenament, i *ii*) manteniment de l'ordre. La primera propietat requereix que el *fitness* extrapoli la inferència de l'usuari en la cerca més enllà dels límits establerts pel propi ordre parcial avaluat. Aquesta propietat garanteix que qualsevol intent d'optimitzar el procés mitjançant aquest *fitness* substitut proporcioni una estimació útil de les possibles avaluacions futures. La segona propietat, manteniment de l'ordre, garanteix un *fitness* sintètic consistent respecte el sistema de selecció per torneig. Llavors, el manteniment de l'ordre garanteix que un *fitness* sintètic serà afinat si és capaç de mantenir l'ordre parcial proporcionat per les decisions de l'usuari. Aquestes assumpcions permeten aplicar tècniques d'optimització senzilles que, malgrat tenir un elevat percentatge d'error, aquest no és important si es manté l'ordre correcte entre les solucions.

Models de *fitness* substitut

Les primeres aproximacions per disposar d'un *fitness* sintètic es van basar en models de proximitat veïnal (*Nearest Neighbour* - NN) (Takagi, 2001). No obstant això aquests models no satisfan les propietats esmentades anteriorment d'extrapolació i d'ordre. Contràriament

al que s'ha explicat, aquests models assumeixen que donat un conjunt de solucions avaluades pels usuaris, la bondat d'una nova solució proposada es pot estimar segons les bondats de les solucions més properes en l'espai segons una mètrica preestablerta (p.ex. correlació lineal del genotip). Altres aproximacions més complexes amitjanen les bondats de les k solucions més properes (k-NN). A continuació s'explica amb detall la motivació que provoca que les aproximacions basades en veïnatge no siguin adequades.



(a) L'extrapolació de *fitness* i el manteniment de l'ordre no es poden garantir sota un model de substitució basat en un model de proximitat veïnal.

(b) L'extrapolació de *fitness* i el manteniment de l'ordre sí es poden garantir sota un model de substitució basat en un model de regressió.

Figura 3.17: Problemes de modelar un *fitness* sintètic mitjançant funcions de veïnatge (a) en comptes d'un model basat en regressió (b) (Llorà *et al.*, 2005*b*).

Primerament cal observar que un model de substitució segons l'heurística de proximitat veïnal no és capaç d'extrapolar la bondat d'una nova solució més enllà dels valors ja avaluats. La figura 3.17(a) exemplifica l'esmentada situació en l'optimització de la funció identitat $f(x) = x$. Qualsevol població finita generada a l'atzar estarà limitada per les dues solucions que adopten la millor i la pitjor bondat avaluada. L'heurística del veí més pròxim produeix una limitació de la capacitat de predicció. A l'optimitzar un *fitness* sintètic, la població final convergirà (en el millor dels casos) al voltant de la millor solució del model de proximitat veïnal. Això succeeix perquè totes les noves solucions possibles, i que són millors que les ja existents (part superior dreta de la línia en la figura 3.17(a)), adopten el mateix valor de bondat, per tant el model no proporciona informació útil sobre aquelles bones solucions que no s'han considerat anteriorment. A més, el model també pot no respectar la

propietat del manteniment de l'ordre parcial. Per exemplificar-ho es pot tornar a fer servir l'exemple de la funció identitat i una heurística de 3-NN (els tres veïns més propers). Per simplificar el problema es consideren els 3 veïns per igual sense cap tipus de ponderació. Al predir la bondat de la solució no considerada $x = 41, f(x) = 41$ es calcula en funció de les tres solucions més properes prèviament avaluades: $s_0 = \{x = 40, f(x) = 40\}$, $s_1 = \{x = 42, f(x) = 42\}$, $s_2 = \{x = 45, f(x) = 45\}$. Segons l'heurística, $\hat{f}(41) = 42'33$. Amb aquest resultat es pot deduir que $f(42) < f(41)$, trencant l'ordre implícit del problema. En canvi, els models de substitució basats en una heurística de regressió sí que respecten les esmentades propietats, tal i com es pot veure a la figura 3.17(b) (Llorà *et al.*, 2005*b*). En el cas dels aiGA s'adopta un regressor basat en màquines de suport vectorial ε-SVM i nucli (*kernel*) lineal. Una descripció detallada de l'ε-SVM es detalla a (Vapnik, 1999; Cristianini i Shawe-Taylor, 2000; Shawe-Taylor i Cristianini, 2004). El *fitness* sintètic basat en ε-SVM fent servir un *kernel* lineal satisfà les propietats d'extrapolació de *fitness* i manteniment de l'ordre. A més, encara que s'obtingui un alt percentatge entre els valors reals i predits (RMSE), la definició del propi model de regressió ε-SVM garanteix l'ordre entre les solucions obtingudes segons un esquema de torneig per selecció. En les proves realitzades per (Llorà *et al.*, 2005*b*) es conclou que per tenir un bon classificador és necessari tenir més de 2ℓ solucions candidates per realitzar la fase d'entrenament on ℓ és el nombre de variables a optimitzar. En canvi, per mantenir l'ordre adequat de les solucions només fan falta $\ell + 2$ exemples.

Si s'assumeix que l'ordre de les variables no afecta el *fitness* avaluat, un *kernel* lineal resulta suficient. Però, en canvi, si l'ordre de les variables en el genotip afectés la qualitat de la síntesi (*linkage*) llavors és necessari emprar un *kernel* polinòmic (Llorà *et al.*, 2005*b*). En el cas dels pesos, no existeix aquest problema, per tant es manté un *kernel* lineal.

Grafs d'ordre parcial

L'ordre parcial de les solucions avaluades es representa com un graf $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$ (tal com s'estipula a (Llorà *et al.*, 2005*b*)). Els vèrtexs de $\mathcal{V}$ representen les solucions mostrades a l'usuari, de la mateixa manera les fletxes de $\mathcal{E}$ representen les avaluacions d'ordre parcial proporcionades per l'usuari. Donades dues solucions candidates $\{s_1, s_2\} \in \mathcal{V}$ l'usuari és capaç d'inferir tres tipus de relacions: *i*) $s_1 > s_2$, *ii*) $s_1 < s_2$, o *iii*) $s_1 = s_2$ (no percep diferències). L'esmentat graf $\mathcal{G}$ es pot transformar en un graf normalitzat $\mathcal{G}'$ sense relacions d'igualtat desfent els empats i propagant, de manera recíproca, les relacions de dominància (Llorà *et al.*, 2005*b*). A continuació s'explica la metodologia amb detall.

Si els tornejos s'ordenen de manera jeràrquica tal i com es presenta en la figura 3.18, es

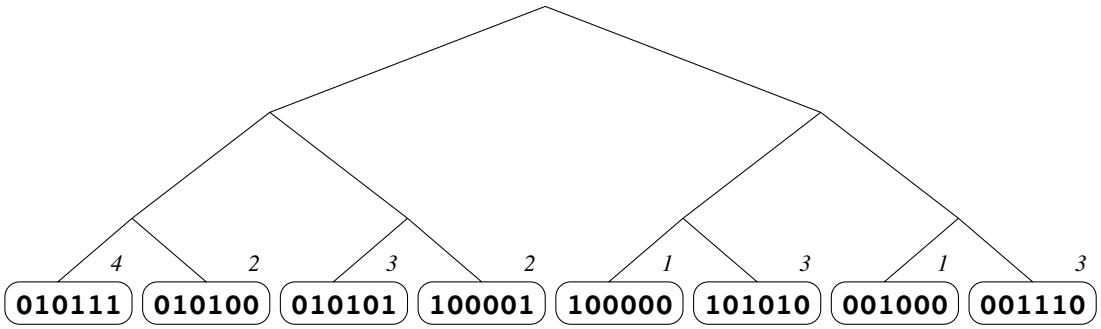
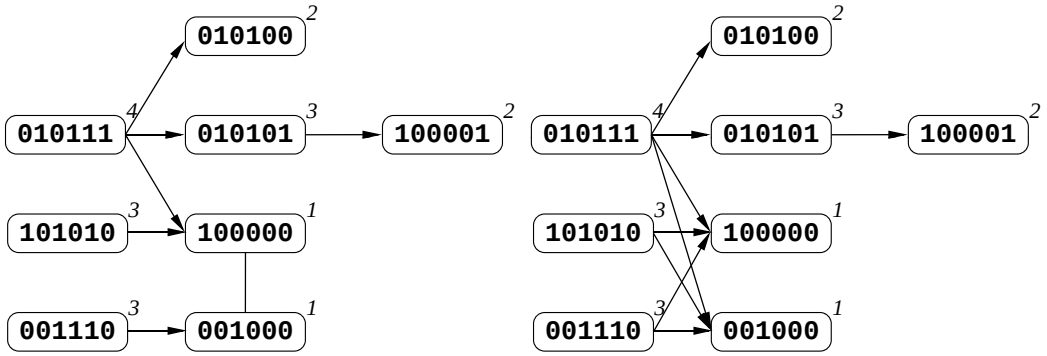


Figura 3.18: Exemple de vuit individus escollits aleatòriament d'una població i assignats en set tornejos diferents. {(010111, 010100), (010101, 100001), (100000, 101010), (001000, 001110), (010111, 010101), (100000, 001000), (010111, 100000)}. Al ser un problema de maximització d'uns (*one-max*). El superíndex de la dreta de cada globus indica la qualitat subjectiva que l'usuari té en ment i que li otorga (Llorà *et al.*, 2005*b*).



(a) *Graf inicial* $\mathcal{G}$ (b) *Graf normalitzat* $\mathcal{G}'$.

Figura 3.19: (a) Graf d'ordre parcial proporcionat per les comparatives de l'usuari a partir dels tornejos de la figura 3.18. (b) Graf amb l'ordre parcial equivalent on les relacions d'igualtat han estat substituïdes per les relacions de dominància dels nodes que conformen l'empat (Llorà *et al.*, 2005*b*).

pot garantir que l'ordre parcial inferit segons les avaluacions de l'usuari produeixi un graf connectat $\mathcal{G}$ (sense subgrafs aïllats). Existeix la possibilitat que l'esmentat graf $\mathcal{G}$ sigui no dirigit (ja que es permeten avaluacions d'igualtat), tanmateix, el mateix procés presentat per Llorà *et al.* (2005*b*) va poder establir formalment i de manera senzilla la transformació del graf $\mathcal{G}$ a un graf dirigit normalitzat $\mathcal{G}'$ (veure figura 3.19) reemplaçant les igualtats (arestes no direccionals) per les corresponents relacions de superioritat o inferioritat (arestes direccionals / fletxes). Aquest procés es realitza mitjançant l'algorisme 3.2.

Algorisme 3.2 Algorisme de normalització de $\mathcal{G}$ a $\mathcal{G}'$ per tal d'eliminar les relacions d'igualtat, on $e(\cdot,\cdot)$ representa la connexió (fletxa) entre dos vertex (Llorà *et al.*, 2005*a*).

procedure normalizeGraph(G=$< \mathcal{V}, \mathcal{E} >$)

1: Crear el conjunt d'empats
$\mathcal{D} = \{\forall e(v_1, v_2) \in \mathcal{E} : \exists e(v_2, v_1) \in \mathcal{E} \Rightarrow e(v_1, v_2) \subseteq \mathcal{D}\}$
2: Crear el conjunt buit $\mathcal{E}_{\mathcal{N}}$ de noves connexions
3: Copiar els camins que arriben als primers items als segons items:
$\mathcal{E}_{\mathcal{N}} \leftarrow \mathcal{E}_{\mathcal{N}} \cup \mathcal{E}_{\mathcal{I}} = \left\{\forall_{e\in\mathcal{D}}(v_1, v_2)\forall v^i | \exists e(v^i, v_1) \in \mathcal{E} \Rightarrow e(v^i, v_2) \subseteq \mathcal{E}_{\mathcal{I}}\right\}$
4: Copiar els camins que arriben als segons items als primers items:
$\mathcal{E}_{\mathcal{N}} \leftarrow \mathcal{E}_{\mathcal{N}} \cup \mathcal{E}_{\mathcal{I}} = \left\{\forall_{e\in\mathcal{D}}(v_1, v_2)\forall v^i | \exists e(v^i, v_2) \in \mathcal{E} \Rightarrow e(v^i, v_1) \subseteq \mathcal{E}_{\mathcal{I}}\right\}$
5: Copiar els camins que surten dels primers items als segons items:
$\mathcal{E}_{\mathcal{N}} \leftarrow \mathcal{E}_{\mathcal{N}} \cup \mathcal{E}_{\mathcal{I}} = \left\{\forall_{e\in\mathcal{D}}(v_1, v_2)\forall v^i | \exists e(v_1, v^i) \in \mathcal{E} \Rightarrow e(v_2, v^i) \subseteq \mathcal{E}_{\mathcal{I}}\right\}$
6: Copiar els camins que surten dels segons items als primers items:
$\mathcal{E}_{\mathcal{N}} \leftarrow \mathcal{E}_{\mathcal{N}} \cup \mathcal{E}_{\mathcal{I}} = \left\{\forall_{e\in\mathcal{D}}(v_1, v_2)\forall v^i | \exists e(v_2, v^i) \in \mathcal{E} \Rightarrow e(v_1, v^i) \subseteq \mathcal{E}_{\mathcal{I}}\right\}$
7: $\mathcal{G}' =< \mathcal{V}', \mathcal{E}' > \leftarrow < \mathcal{V}, \emptyset >$
8: $\mathcal{E}' \leftarrow \mathcal{E} \cup \mathcal{E}_{\mathcal{N}}$
9: $\mathcal{E}' \leftarrow \mathcal{E}' - \mathcal{D} - \mathcal{D}_{\mathcal{N}} = \{\forall e(v_1, v_2) \in \mathcal{D} : e(v_2, v_1) \subseteq \mathcal{D}_{\mathcal{N}}\}$
10: *normalizeGraph* $\leftarrow \mathcal{G}'$

Donat un graf d'ordre parcial normalitzat $\mathcal{G}'$, es pot crear un *fitness* sintètic seguint el concepte de dominància de Pareto (Pareto, 1896). El fet d'aplicar el criteri de dominància de Pareto no és nou en la computació evolutiva ja que és una tècnica que ha donat molt bons resultats per l'adaptació dels GA a problemes d'optimització multiobjectiu (Coello-Coello, December, 1998; Deb *et al.*, 2000). Es pot establir un ordre de bondat en les solucions a partir d'una heurística basada en dues mesures de dominància, δ and ϕ. $\delta(v)$ és el nombre de vèrtexs diferents que es troben en tots els camins que parteixen del vèrtex v i $\phi(v)$ és el nombre de vèrtexs diferents que es troben en tots els camins que arriben al vèrtex v. La taula 3.1 mostra $\delta(v)$ i $\phi(v)$ donat el graf de la figura 3.19(b). Llavors, segons l'heurística,

v	$f(v)$	$r(v)$	$\delta(v)$	$\phi(v)$	$\hat{f}(v)$	$\hat{r}(v)$
010111	4	1	5	0	5	1
010100	2	3	0	1	-1	5
010101	3	2	1	1	0	4
100001	2	3	0	2	-2	6
100000	1	4	0	3	-3	7.5
101010	3	2	2	0	2	2.5
001000	1	4	0	3	-3	7.5
001110	3	2	2	0	2	2.5

Taula 3.1: Estimació de la classificació global basada en la mesura de dominància emprant l'ordre parcial mostrat a la figura 3.19(b). Pel càlcul de $\hat{r}(v)$ veure selecció basada en *ranking* de l'apartat 3.2.3

el valor de *fitness* estimat donada una solució v és pot calcular com $\hat{f}(v) = \delta(v) - \phi(v)$. Intuïtivament, quantes més solucions són dominades pel vèrtex v (*és superior*), el valor de *fitness* és millor. En canvi, quantes més solucions dominen el vertex v (*és inferior*) pitjor és el seu valor de *fitness*. La classificació final de manera global $\hat{r}(v)$ s'obté ordenant els vèrtexs $v \in \mathcal{V}$ mitjançant $\hat{f}(v)$, tal i com es mostra en la taula 3.1. Aquest estimador global $\hat{r}(v)$ es pot fer servir per entrenar un regressor basat en màquines de suport vectorial ε-SVM (Vapnik, 1999; Cristianini i Shawe-Taylor, 2000; Shawe-Taylor i Cristianini, 2004) i així obtenir el *fitness* sintètic que permet una cerca activa de l'espai de possibles solucions: trobant les solucions òptimes segons aquest *fitness* sintètic permet escollir propostes intel·ligents (*educated guesses*) per a les subseqüents iteracions amb l'usuari.

3.4.4 Optimització de la cerca activa: Algorisme Genètic Compacte (cGA)

Un dels canvis més importants que realitza Llorà *et al.* (2005*b*) en la proposta de l'aiGA respecte l'iGA és passar d'un mètode de cerca basat purament en població a un mètode de cerca basat en estimació i distribució de dades (EDA), que es comporta igual que un GA clàssic (Harik *et al.*, 1999) amb selecció mitjançant *steady-state* (veure l'apartat 3.2.3) i creuament uniforme (veure apartat 3.2.3). Les propietats de creuament uniforme i *steady-state* es troben tanmateix en l'iGA clàssic (Takagi, 2001) que realitza una selecció per torneig $s = 2$.

L'algorisme genètic compacte (cGA) (Harik *et al.*, 1999) és una optimització dels EDA (l'EDA típic és l'algorisme de *Population Based Incremental Learning* — PBIL). El cGA basa el seu funcionament en un vector de distribucions de probabilitat en comptes d'una població. Per tant, es treballa amb estadístiques de la població i no amb els propis individus. El cGA

processa cada gen de manera independent i per tant requereix menys memòria que un algorisme genètic simple. Per tant, es realitza una ràpida estimació del problema.

El funcionament del cGA es detalla en l'algorisme 3.3. El funcionament de l'algorisme es basa en generar un vector de probabilitats neutre p de mida l (igual probabilitat de generar tant 0s com 1s en la seva representació binària) i actualitzar-lo en funció del comportament mitjançant torneig de parelles de solucions aleatòries generades a partir de l'esmentat vector p. El procés es repeteix fins que la probabilitat és 1 o 0 per a totes les probabilitats del vector.

Algorisme 3.3 Pseudocodi de l'algorisme genètic compacte (cGA) (Harik *et al.*, 1999).

procedure $cGA(l)$

1: **for** $i = 0$ to l **do**
2: $p[i] \leftarrow 0.5$
3: **end for**
4: $a \leftarrow genera(p)$
5: $b \leftarrow genera(p)$
6: $millor, pitjor \leftarrow avalua(a, b)$
7: **for** $i = 0$ to l **do**
8: **if** $millor[i] \neq pitjor[i]$ **then**
9: **if** $millor[i] = 1$ **then**
10: $p[i] \leftarrow p[i] + 1/n$
11: **else**
12: $p[i] \leftarrow p[i] - 1/n$
13: **end if**
14: **end if**
15: **end for**
16: **for** $i = 0$ to l **do**
17: **if** $(p[i] > 0) \cap (p[i] < 1)$ **then**
18: $goto(4)$
19: **end if**
20: **end for**
21: $retorna(p)$

Les dues principals diferències del cGA respecte el PBIL (Baluja i Caruana, 1995) són que *i*) pot executar un algorisme genètic sense diposar de tota població i *ii*) reduint els requeriments de memòria del GA.

L'increment d'actualització d'un cGA té una mida constant de $1/n$. Mentre que l'algorisme genètic simple necessita emmagatzemar n bits per a cada posició del gen, el cGA només necessita mantenir la proporció d'uns (respecte la proporció de zeros), un conjunt finit de n pot ser emmagatzemat amb $\log_2(n)$ bits.

3.4.5 Algorismes genètics interactius actius

Algorisme 3.4 Descripció algorítmica de l'aiGA (Llorà *et al.*, 2005*b*).

procedure $aiGA()$

1: Crear un graf buit $\mathcal{G}$ dirigit
2: Crear 2^h solucions inicials aleatòries (conjunt $\mathcal{V}$)
3: Crear el conjunt de torneig jeràrquic $\mathcal{T}$ fent servir les solucions disponibles a $\mathcal{V}$
4: Presentar els tornejos de $\mathcal{T}$ a l'usuari i actualitzar l'ordre parcial a $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$
5: Estimar $\hat{r}(v)$ per cada $v \in \mathcal{V}$
6: Entrenar el *fitness* sintètic substitut ε-SVM basat en $\mathcal{V}$ i $\hat{r}(v)$
7: Optimitzar el *fitness* sintètic substitut ε-SVM fent servir el cGA.
8: Crear un conjunt $\mathcal{V}'$ amb 2^{h-1} solucions diferents on $V \cap V' = \emptyset$ mostrejant el model probabilístic evolucionat pel cGA.
9: Crear un conjunt de torneig jeràrquic $\mathcal{T}'$ amb $2^h - 1$ tornejos fent servir 2^{h-1} solucions en $\mathcal{V}$ i 2^{h-1} solucions en $\mathcal{V}'$
10: $\mathcal{V} \Leftarrow \mathcal{V} \cup \mathcal{V}'$
11: $\mathcal{T} \Leftarrow \mathcal{T} \cup \mathcal{T}'$
12: Saltar a 4 mentres no convergeixi

Repassant els apartats anteriors es pot concloure que la interacció entre l'usuari i el procés evolutiu genera informació que es pot processar més enllà del propi procés evolutiu. Aquesta anàlisi persegueix dos objectius: *i*) realitzar mineria de dades sobre les avaluacions de l'usuari i *ii*) entendre els patrons de comportament (de manera similar a l'aprenentatge actiu). No obstant això, els processos IEC han d'assumir que el coneixement proporcionats pels usuaris és limitat i independent en relació al coneixement general del problema: el mateix problema pot tenir solucions diferents depenent dels usuaris.

Tal com s'ha vist, el conjunt d'ordenacions parcials de les solucions es pot transformar en un ordre complet induït que es pot transformar en un graf dirigit el qual representa la preferència de l'usuari. Aquest model de graf pot servir per entrenar un sistema basat en un algorisme d'aprenentatge artificial combinat amb un algorisme d'optimització (cGA)

que aporta propostes intel·ligents (*educated guesses*) a l'usuari anàlogament als mètodes de cerca actius (veure apartat 3.4.1). D'aquesta manera s'eviten interaccions (avaluacions) absurdes o redundants i s'aconsegueix disminuir la fatiga de l'usuari avaluador durant el transcurs del procés (Llorà *et al.*, 2005*b*). L'algorisme general d'un aiGA es detalla en l'algorisme 3.4.

3.4.6 Adaptació de l'aiGA al problema de l'ajust de pesos en CTP-SU

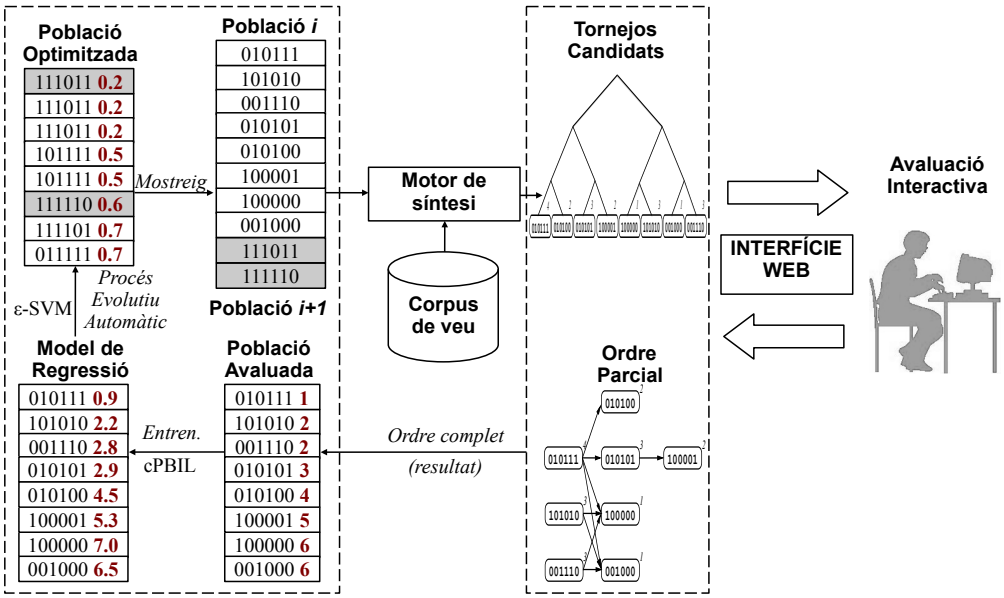


Figura 3.20: Flux d'execució de la metodologia d'ajust de pesos mitjançant aiGA per sistemes de conversió de text a parla basats en selecció d'unitats.

A (Alías, 2006) s'estudia un principi de viabilitat de l'idoneïtat de l'aiGA proposat per (Llorà *et al.*, 2005*b*) al problema de l'ajust de pesos. A la figura 3.20 es presenta l'adaptació realitzada (diagrama de blocs) per optimitzar els pesos de la funció de cost (equació (2.4)), considerant les especificitats del problema. Com es pot observar, en el procés d'optimització de la població es substitueix l'algorisme genètic compacte (cGA) utilitzat a (Llorà *et al.*, 2005*b*) per un altre algorisme de la família dels algorismes PBIL: aquest canvi permet treballar amb gens de valor real i contínu (pesos). L'algorisme s'anomena PBIL continu (cPBIL) (Sebag i Ducoulombier, 1998). Així mateix, el sistema de regressió ε-SVM necessita d'una adaptació al problema. A continuació es descriuen els canvis en el ε-SVM així com els trets fonamentals del cPBIL.

Adaptació del ε-SVM a valors continus

Com s'ha comentat, l'ordre complet induït obtingut a partir del graf d'ordenació parcial que modela les preferències de l'usuari permet generalitzar un model per obtenir una funció de *fitness* sintètic que englobi tot l'espai de cerca. Aquest modelat es realitza mitjançant un sistema de regressió ε-SVM, capaç d'avaluar les solucions d'acord amb la resposta de l'usuari fins a aquest moment. Aquest necessita ajustar-se per funcionar adequadament sobre un conjunt de valors reals. En aquest sentit, es va emprar la llibreria SVM de la National Taiwan University (Chang i Lin, 2001). Aquesta adaptació es va realitzar mitjançant les eines d'adaptació que aporta la pròpia implementació i forma part del treball preliminar realitzat amb anterioritat a aquesta tesi doctoral – veure (Formiga, 2005) per a una descripció detallada de tot el procés.

Tal com s'ha dit (apartat 3.4.3), el ε-SVM fixa el límit inferior del nombre de tornejos que han de ser avaluats per l'usuari. En les proves realitzades per Llorà *et al.* (2005*b*), es demostra que per entrenar el sistema de regressió ε-SVM amb un *kernel* lineal sobre un espai de dimensió ℓ cal disposar de $\ell + 2$ exemples (individus amb el seu *fitness* – veure taula 3.19(b)). No obstant això, en l'adaptació, donada la complexitat del problema d'optimització, es va augmentar el nombre d'individus avaluats per l'usuari al doble ($\ell = 16$ – valor fixat empíricament – Formiga (2005)) per a disposar de suficient informació per obtenir una estimació fiable del *fitness* subjectiu.

Population Based Incremental Learning en espais continus (cPBIL)

Seguint el treball de Sebag i Ducoulombier (1998), l'algorisme cPBIL emprat en l'adaptació s'estructura en els següents passos (veure figura 3.21):

1. *Inicialització:* Anàlogament a un algorisme genètic simple, que inicia aleatòriament la població, en un cPBIL es modela el valor de cada gen de l'individu mitjançant una distribució normal $N(X, \sigma)$, on $X_i = 0.5$ (Sebag i Ducoulombier, 1998) y $\sigma_i = 0.1$ (valor fixat experimentalment), per a $1 \leq i \leq \ell$ gens de cada individu – en aquest cas el nombre de pesos a ajustar ($\ell = param\_t + param\_c$, segons les equacions 2.2 i 2.3).

2. *Mostreig del model:* Cada nou individu es genera a partir del conjunt de valors aleatoris obtinguts del mostreig de la distribució normal que representa cada un dels gens (pesos). Es generaran tants individus aleatoris com la grandària de la població n indiqui.

3. *Valoració:* Es calcula el *fitness* o mesura de qualitat dels individus – obtinguda, en aquest cas, a partir del resultat de l'aplicació del regressor ε-SVM – i s'ordena la població segons el resultat obtingut (com més gran sigui el valor de *fitness* de l'individu, millor estarà adaptat al criteri de l'usuari).

4. *Selecció:* Com en els algoritmes genètics tradicionals, cPBIL disposa d'un esquema de selecció encarregat d'escollir els individus que actualitzaran les mitjanes de les distribucions normals representades en els gens dels individus.En aquest cas, seguint el que indica per (Sebag i Ducoulombier, 1998), es seleccionen els K millors individus (exemples positius) amb el pitjor (exemple negatiu) de la població per actualitzar el model probabilístic.

5. *Actualització del model probabilístic:* D'una banda, s'actualitza la mitjana de les distribucions dels gens dels individus, tenint en compte els dos millors individus costat del pitjor de la població, segons l'equació 3.3 (Sebag i Ducoulombier, 1998).

$$X_i^{t+1} = (1-\alpha)\cdot X_i^t + \alpha\cdot(X^{best_1} + X^{best_2} - X^{worst}), \tag{3.3}$$

   on t indica la iteració actual i α ($0 \leq \alpha \leq 1$) representa el factor d'aprenentatge (*learning rate*, en anglès) que determina la velocitat de la convergència del vector de probabilitats.

   A continuació, s'actualitza la desviació de les distribucions dels gens a partir de la variància dels K millors individus, segons l'equació (3.4) – veure (Sebag i Ducoulombier, 1998) per a una justificació de l'estratègia utilitzada.

$$\sigma_i^{t+1} = (1-\alpha)\sigma_i^t + \alpha\sqrt{\frac{\sum_{j=1}^{K}(X_i^j - \hat{X}_i^K)}{K}} \tag{3.4}$$

   on i indica el gen, j l'individu i $\hat{X}_i^K$ representa la mitjana del gen i dels K millors fills $\{X^1, \ldots, X^K\}$.

6. Repetir els passos 2 a 5 fins assolir el criteri de convergència.

En aquest treball s'empra $K = n/5$, seguint les indicacions de (Sebag i Ducoulombier, 1998), $\alpha = 1/n$ segons (Llorà *et al.*, 2005*b*) i es fixa experimentalment el criteri de parada a $\sigma = 0.01$, pel fet que es treballa amb pesos amb resolució de dos decimals. A continuació s'aprofundeix en la precisió de les dades.

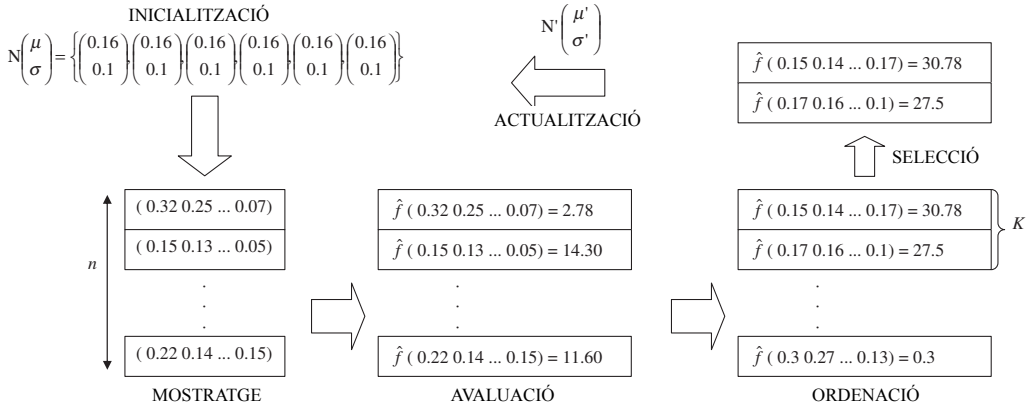


Figura 3.21: Diagrama del funcionament de l'algorisme genètic cPBIL basat en la representació de la població mitjançant distribucions probabilístiques $N(\mu, \sigma)$, en aquest cas amb 6 gens per individu (Alías, 2006).

Normalització de les dades i precisió de treball

Un cop fixats i adaptats el sistema de regressió (ε-SVM) i el procés evolutiu (cPBIL), s'ha de considerar la restricció que la suma dels pesos ha de ser 1 ($\sum_i^N w_i = 1$). En aquest sentit resulta fonamental decidir on es realitza la normalització de les dades i amb quina precisió es treballarà en tot moment. Es decideix treballar amb una precisió de dos decimals ja que, en les proves realitzades amb anterioritat (Alías *et al.*, 2003) demostra aportar bona sensibilitat a la selecció d'unitats (Alías, 2006). Tal com es mostra a la taula 3.2, sobre el conjunt de frases utilitzat en les proves subjectives prèvies, aquesta resolució aconsegueix una ràtio de 0.677 entre el nombre de pesos (genotips) i el nombre de frases diferents sintetitzades a partir dels mateixos (fenotips) – es computen com diferents totes aquelles frases que contenen, com a mínim una unitat diferent a la resta de frases. Si es treballa amb més precisió, d'una banda, les diferències entre els resultats sintètics resulten menys perceptibles, i d'altra, s'augmenta el soroll del procés evolutiu, ja que l'aiGA és incapaç de reaccionar de forma diferent (escollint solucions diferents) davant de canvis tan petits en els valors dels pesos. D'altra banda, treballar amb menor precisió (1 decimal) provocaria una ràtio – entre genotips i fenotips – superior a 1, de manera que el escombrat d'unitats del corpus seria deficient.

Pel que fa a la normalització dels individus, en l'esquema del propi aiGA, resulta necessari mantenir les distribucions sense normalitzar els valors dels pesos durant el procés evolutiu per permetre la convergència cPBIL, pel fet que aquest representa la població com

Frase	Candidates	Pesos	Ratio
"De la seva selva"	82	132	62.12%
"Fusta de Birmània"	82	118	69.49%
"Grans extensions"	81	109	74.31%
"I els han venut"	75	125	60.00%
Total	406	602	67.44%

Taula 3.2: Relació entre fenotips i genotips a (Alías, 2006). La columna ràtio indica la relació entre el nombre de frases candidates i el nombre de vectors de pesos utilitzats per obtenir-les.

distribucions normals $N(\mu, \sigma)$. Així doncs, la normalització es realitza una vegada ha finalitzat l'ajust interactiu dels pesos. Si això no fos així, durant la implementació, s'ha pogut observar que la normalització intermèdia de les dades provoca inestabilitats en el procés evolutiu degut a que els valors μ i σ del vector de pesos oscil·len provocant que l'aiGA es comporti erràticament.

3.4.7 Mesura de la consistència de l'usuari

En treballs previs (Alías *et al.*, 2003) es va observar les propietats que defineixen l'escenari del problema de l'ajust subjectiu de pesos. Una d'aquestes propietats és la diferència entre el domini del conjunt de genotips i de fenotips. Per exemple, de 15 possibles individus (configuracions de pesos, o genotips), només es sintetitzen 4 o 5 solucions diferents (fitxers de so, o fenotips). Tal com s'ha explicat a l'apartat 3.4.3, l'ordre parcial entre les solucions es modela mitjançant un graf $\mathcal{G}' =< \mathcal{V}, \mathcal{E} >$ on els possibles individus son els vèrtexs $\mathcal{V}$ i les relacions de qualitat entre les solucions són les fletxes del graf $\mathcal{E}$. Aquest fet, dóna a entendre que els tornejos poden esdevenir molt sorollosos i que, a la vegada, això provoca que es puguin produir molts empats o contradiccions durant les comparatives. En la figura 3.22 es pot observar un cicle entre tres individus ({2,4,5}). El cicle es pot entendre com un conjunt de nodes que comparteixen les mateixes dominàncies respecte la resta de nodes en el graf (zona d'equidominància entre els individus). Una possible hipòtesi d'aquest resultat és que els tres individus del cicle tinguin el mateix fenotip i, per tant, sigui correcte que comparteixin dominància (i ordre parcial) en la població. Contràriament, a l'hora d'analitzar aquests grafs dirigits, també es poden entendre aquests cicles com a inconsistències de l'usuari, el qual pren decisions d'ordre parcial contradictòries que provoquen l'aparició de cicles en el graf d'ordenació global.

En aquest sentit, es va realitzar un primer estudi de viabilitat del mètode *off-line* tot explorant els registres d'avaluacions (*logs*) de les proves d'ajust realitzades amb iGA (veu-

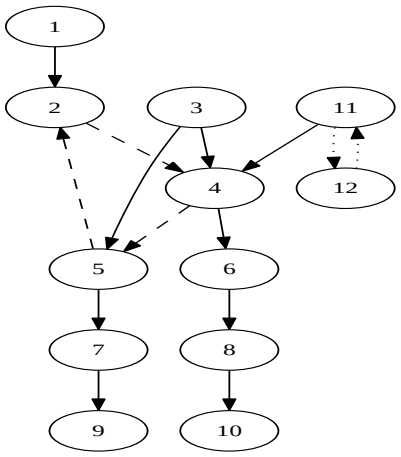


Figura 3.22: Exemple d'un cicle amb equidominància dins una població.

re 3.3.3). En aquesta anàlisi, es va observar que en els grafs que se'n derivaven podien contenir molts cicles i la gran majoria no compartien fenotip.

La hipòtesi de disseny que es va assumir llavors fou, tot seguint el raonament descrit a (Takagi, 2001), que l'usuari no pugués mantenir el seu criteri durant moltes iteracions. Llavors, el fet que sigui altament probable l'existència de cicles en el graf no permet seguir desenvolupant i adaptant el procés sense madurar una mica a quin tipus de solucions conduïen aquests cicles, com influïen a la solució final del procés i com es podia millorar la qualitat d'aquesta solució.

Llavors, aquesta circumstància es pot representar de manera formal mitjançant la següent propietat interesant: donat una graf d'ordre parcial normalitzat $\mathcal{G}'$, si un vèrtex v apareix més d'una vegada en un camí quan es calcula $\delta(v)$ o $\phi(v)$, llavors es pot afirmar que el graf conté un cicle (Alías *et al.*, 2006*a*).

Implementació de la detecció de cicles

Una vegada detectats els nodes que formen part d'un cicle, resulta necessari definir algun procés encarregat d'obtenir un ordre robust de les solucions evitant l'aparició d'inconsistències en la ordenació. L'objectiu és disposar d'una classificació ordenada de les solucions per a entrenar el regressor lineal usat, sense trencar físicament els cicles. Disposar d'un algorisme amb capacitat de detectar cicles és de vital importància per mesurar la consistència de l'usuari durant el procés d'avaluació.

Segons aquest criteri, a (Llorà *et al.*, 2005*a*) es proposa l'implementació d'algorismes

Algorisme 3.5 Algorisme de detecció de cicles dins $\mathcal{G}'$.

procedure *cycleDetection*$(\mathcal{G}')$

1: Crear el conjunt buit de $\mathcal{C}$ cicles, $\mathcal{V}_\mathcal{T}$ de vèrtex visitats.
2: Extraure el primer vertex $v^i \in \mathcal{V} \mid v^i \notin \mathcal{V}_\mathcal{T}$
3: Crear el conjunt$\mathcal{V}_\mathcal{N} = \{\forall\, v \in \mathcal{V}_\mathcal{N} : (v \neq v^i) \cap (v \in \mathcal{G}')\}$
4: Crear el conjunt de cicles $\mathcal{C}_\mathcal{I} = \{\forall v \in \mathcal{V}_\mathcal{N}\ \forall e(v^i, v) \in \mathcal{E}' : cycleExplorer(\{v^i\}, v, \mathcal{G}') \subseteq \mathcal{C}_\mathcal{I}\}$
5: Eliminar les parts no cícliques i treure cicles repetits (degut al multicicle) $\forall c \in \mathcal{C}_\mathcal{I}$
6: Ordenar els cicles considerant el vèrtex més antic com a primer i últim del camí a $\forall c \in \mathcal{C}_\mathcal{I}$
7: $\mathcal{V}_\mathcal{T} \leftarrow \mathcal{V}_\mathcal{T} \cup \{v^i\}$
8: $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}_\mathcal{I}$
9: Tornar al pas 2 mentre $\forall v^i \in \mathcal{G}'$: $v^i \notin \mathcal{V}_\mathcal{T}$
10: $cycleDetection \leftarrow \mathcal{C}$

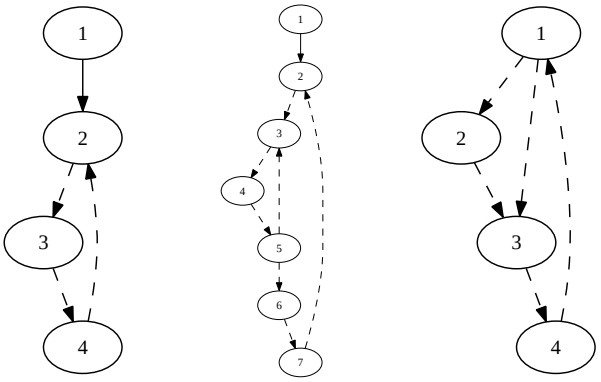
Algorisme 3.6 Exploració de tots els camins que surten del vèrtex v a $\mathcal{V}_\mathcal{I}$, on $e(\cdot,\cdot)$ representa la connexió (fletxa) entre dos vèrtexs.

procedure *cycleExplorer*$(\mathcal{V}_\mathcal{I}, v, \mathcal{G}' =< \mathcal{V}', \mathcal{E}' >)$

1: $\mathcal{V}_\mathcal{I} \Leftarrow \mathcal{V}_\mathcal{I} \cup v$
2: Crear el conjunt $\mathcal{R} = \{\forall v^i \in \mathcal{V}_\mathcal{I}\ \forall e(v, v^i) \in \mathcal{E}' : e(v, v^i) \subseteq \mathcal{R}\}$
3: $(\mathcal{R} \neq \emptyset) \Rightarrow return(\mathcal{R})$
4: Crear el conjunt $\mathcal{C}_\mathcal{I} = \{\forall v^i \in (\mathcal{V} - \{v\}) \forall e(v, v^i) \in \mathcal{E}' : cycleExporer(\mathcal{V}_\mathcal{I}, v^i, \mathcal{G}') \subseteq \mathcal{C}_\mathcal{I}\}$
5: $return(\mathcal{C}_\mathcal{I})$

(algorismes 3.5 i 3.6) que realitzen la detecció de cicles en el graf basant-se en les avaluacions que infereix l'usuari. El desenvolupament d'aquests algorismes formen part de la col·laboració en la recerca realitzada abans d'aquesta tesi doctoral (Formiga, 2005). Aquests algorismes persegueixen l'objectiu de mantenir el mateix criteri a l'hora d'identificar els cicles i, per tant, evitar les redundàncies que aquests provoquen a l'hora d'explorar camins. Els algorismes funcionen de la manera següent:

Per cada vèrtex v del graf $\mathcal{G}' =< \mathcal{V}', \mathcal{E}' >$, els algorismes exploren la relació amb els altres vèrtexs $\mathcal{V}^{\mathcal{N}}$ que encara no s'han processat. Aquesta relació s'analitza a través d'un conjunt incremental de vèrtexs visitats en el transcurs d'un camí. Un cop el conjunt de camins cíclics detectats queda definit, les parts no cícliques dels diversos camins es filtren per evitar l'ambigüitat de tenir subcicles. Finalment, aquest conjunt de camins cíclics detectats és el que es retorna com a solució de l'algorisme.



(a) Cicles. (b) Subcicles. (c) Multicicles.

Figura 3.23: Exemples de diferents tipologies de (a) cicles, (b) subcicles i (c) multicicles.

A tal efecte, un cop s'han detectat els camins cíclics mitjançant l'algorisme 3.5, les parts no cícliques dels camins, que són les solucions correctament ordenades, es filtren per a evitar l'ambigüetat dels subcicles. Un dels objectius d'aquest procés és tenir en compte els subcicles que poden aparèixer dins d'un macro-cicle d'ordre superior, considerant-los com a cicles diferents. Un exemple el podem trobar en els tres cicles mostrats a la figura 3.23(c), on es pot observar com el tercer graf conté un subcicle, que cal analitzar com a dos cicles.

Eliminació de cicles

A nivell de qualitat de *software* l'aparició d'un cicle pot provocar que certs algorismes caiguin en un bloqueig actiu en intentar processar el graf. Alguns exemples d'aquests bloquejos podrien produir-se en buscar la dominància d'un node, desfer els empats o buscar el camí més llarg dins del graf. Per tant, s'ha d'eliminar les parts cícliques del graf per a ser poder tractades pels altres mòduls de l'aiGA.

Per dur a terme l'eliminació de cicles, en el treball desenvolupat a (Formiga, 2005) es defineixen dues heurístiques:

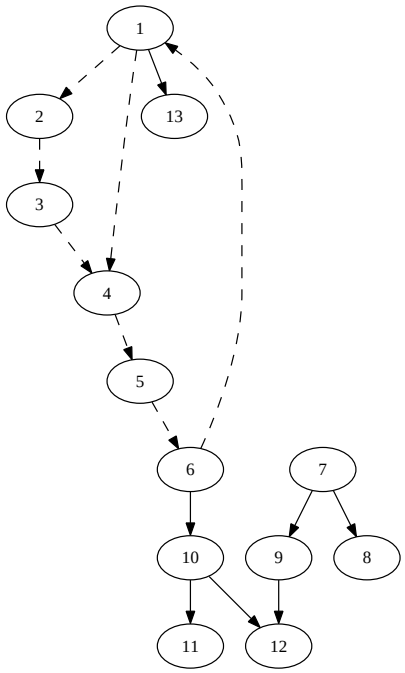


Figura 3.24: Exemple d'un cas de cicles algorítmicament complex, on se suposa la mateixa ponderació per a cada connexió (fletxa).

i) Primerament s'etiqueta cada fletxa, que és la relació de preferència entre solucions, amb el nombre de vegades que aquesta relació ha estat ratificada per l'usuari (quantitat final de vots), trencant el cicle per la connexió de menor puntuació. El fet d'emprar un torneig jeràrquic durant les generacions, permet que un mateix torneig sigui avaluat diferents vegades. D'aquesta manera, es pot determinar una puntuació de solidesa per cada fletxa, definint com el nombre de vegades que l'usuari ha votat en aquest sentit la comparativa de solucions presentada. Per exemple, donat el graf cíclic $\{3 \rightarrow 4 \rightarrow 5 \rightarrow 3\}$, amb les fletxes etiquetades segons les votacions

$\{3 \rightarrow 4\} : 2, \{4 \rightarrow 5\} : 1$ i $\{5 \rightarrow 3\} : 2$, el cicle es trencaria per la connexió $\{4 \rightarrow 5\}$, que, en aquest cas, és la de menor puntuació, la menys fiable. Aquesta heurística va ser estudiada en més profunditat, sota un punt de vista probabilístic per Llorà *et al.* (2008)

ii) En el cas que totes les connexions presentin la mateixa puntuació, és necessari un criteri addicional per decidir on trencar el cicle. En aquest sentit s'estableix una nova heurística que considera la dominància de cada vèrtex eliminant les fletxes que conformen el cicle. Llavors es suprimeix la fletxa que va des del vèrtex amb menor dominància fins al vèrtex amb major dominància. En aquest sentit és el cicle en el punt més dèbil en termes de dominància. Finalment, i com a últim criteri usat, si totes dues heurístiques no permeten trencar el cicle, s'elimina la fletxa del cicle més recent en la població (més jove generacionalment parlant). Una vegada eliminada la fletxa segons els criteris descrits, es procedeix a determinar la dominància del vèrtex de la mateixa manera que s'ha descrit anteriorment, mitjançant les heurístiques $\delta(v)$ i $\phi(v)$. En la figura 3.24 s'observa com la fletxa 6→1 és la que ocasiona l'aparició del cicle ja que trenca tot l'ordre preestablert del graf. Aquest graf amb cicles pot haver-se generat a conseqüència de que tota la zona del cicle tingui el mateix fenotip o bé degut a que l'usuari no hagi estat coherent en alguna avaluació. Per tal d'eliminar aquest cicle, considerant que totes les fletxes tenen la mateixa puntuació, l'algorisme que es proposa a (Formiga, 2005) funcionaria correctament ja que eliminaria la fletxa que va des d'un vèrtex amb menor dominància a un vèrtex amb major dominància.

Tanmateix, cal puntualitzar el fet que els cicles del graf només s'eliminen per a poder obtenir una ordenació consistent dels nodes i així, poder generar el model de *fitness* sintètic que guiï la generació de noves solucions mitjançant l'algorisme cPBIL. Per tant, els cicles continuen representats en el graf fins que l'usuari trenca el cicle repetint alguna de les comparacions durant el procés evolutiu. Quan l'usuari es contradeix invertint el sentit d'una fletxa durant el transcurs del temps es considera que la fletxa més recent com aquella vàlida, al considerar que l'usuari ha madurat el criteri a base d'avaluar fenotips. El fet que un usuari pugui invertir o ratificar el sentit d'una fletxa, ofereix una validació contínua del criteri del propi usuari.

Mesura de consistència d'usuari

La presència de cicles en el graf representa una interacció inconsistent de l'usuari en el procés d'ajust interactiu (Alías *et al.*, 2006*a*). Llavors, disposar d'un graf dirigit permet

identificar la consistència de les avaluacions de l'usuari. Aquesta propietat és la base de la mètrica de consistència que es proposa a (Alías *et al.*, 2006*a*). Un usuari serà consistent en l'instant de temps t si no es troba cap cicle en el seu graf corresponent d'ordre parcial $\mathcal{G}'$ normalitzat.

Per tal de calcular l'esmentada mètrica són necessàries dues components: un algorisme, tal com s'ha explicat, amb capacitat de detectar cicles donat un graf $\mathcal{G}'$ en l'instant de temps t ($\mathcal{G}'^t$), i una heurística per quantificar el grau d'inconsistència que provoca cada cicle detectat, que es pot definir com (Alías *et al.*, 2006*a*):

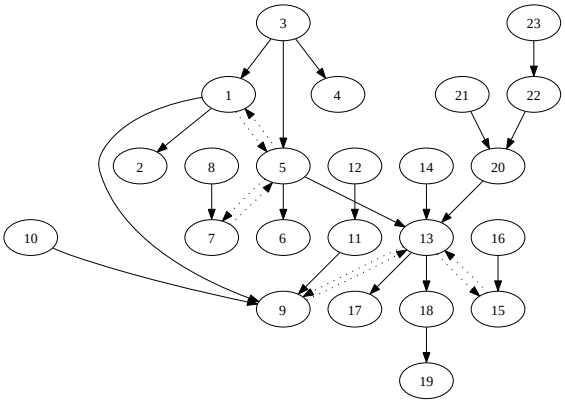
$$\kappa\left(\mathcal{G}'^t, w\right) = 1 - \left(\frac{1}{\left|\mathcal{V}'^t\right|} \cdot \sum_{v \in \chi(\mathcal{G}'^t)} w_v\right)^{\alpha} \tag{3.5}$$

on $\left|\mathcal{V}'^t\right|$ és el nombre total de vèrtexs dins $\mathcal{G}'$ durant l'instant t, w_v la ponderació de cada vèrtex v (no s'ha de confondre amb els pesos de la funció de cost de selecció), $\chi\left(\mathcal{G}'^t\right)$ els vèrtexs totals que formen part de qualsevol cicle existent $\mathcal{G}'^t$, i α un factor d'escala global igual o major que 1. En l'aproximació d'aquesta dissertació doctoral es considera que, $w_v = 1, \forall v \in \mathcal{V}'^t$ and $\alpha = 1$.

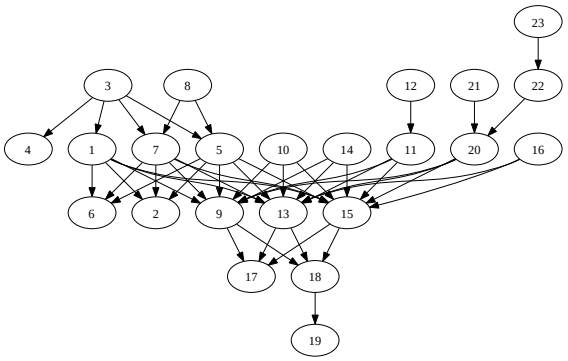
A les figures 3.25, 3.26 i 3.27, es presenten diferents exemples del càlcul de la mesura κ per tres iteracions d'un procés d'interacció amb l'usuari. En elles es representen els grafs d'ordenació parcial, els valors de *fitness* subjectiu i el rànquing calculat, juntament amb el valor de la consistència de l'usuari en l'instant t, $\kappa(\mathcal{G}'^t, w)$. En aquests exemples es pot observar l'impacte de la inconsistència de l'usuari en la presa de decisions sobre la mesura de consistència. A la figura 3.25 es mostra el càlcul en un model sense cicles. A la figura 3.26 s'observa com la fletxa $\{9 \rightarrow 1\}$ és la que ocasiona l'aparició del cicle, és a dir, l'usuari ha preferit la solució 9 abans que la 1, després d'haver preferit la 13 a la 9, i la 1 a la 13, trencant, doncs, l'ordre establert en el graf. La situació es repeteix en la figura 3.27, al connectar-se $\{16 \rightarrow 1\}$,$\{11 \rightarrow 1\}$ i $\{13 \rightarrow 9\}$ i considerar com empat $\{13 \leftrightarrow 15\}$, en aquest cas es fa molt difícil considerar aquests resultats com a bons. Tot i així, es pot veure l'eficiència de l'algorisme d'eliminació de cicles observant la diferència entre les figures 3.27(a) i 3.27(b).

3.4.8 Experiments i resultats

Per tal de validar l'aiGA, es va realitzar una experimentació que perseguia dos objectius: *i*) explorar la millora de consistència i eficiència de les avaluacions de l'usuari en l'entorn d'ajust de pesos per CTP-SU i *ii*) obtenir un nou conjunt de pesos perceptiu de manera



(a) Graf d'ordenació parcial.



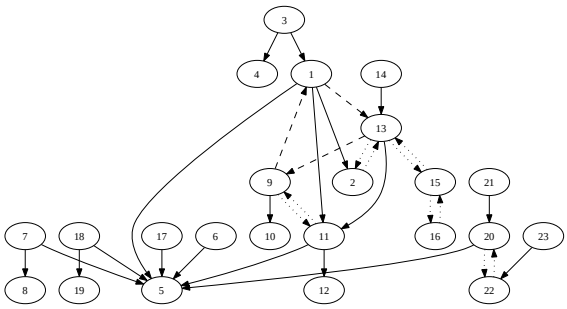
(b) Graf d'ordenació parcial havent normalitzat els empats.

Individu	**DUR.T**	**ENE.C**	**ENE.T**	**MFC.C**	**PIT.C**	**PIT.T**	$\hat{r(v)}$	ε-**SVM** $\hat{r}(v)$
3	0.27	0.26	0.12	0.11	0.06	0.19	23	(22.19)
8	0.19	0.3	0.23	0.09	0.05	0.13	22	(21.14)
23	0.18	0.18	0.19	0.13	0.04	0.28	21	(20.19)
21	0.33	0.1	0.13	0.27	0.1	0.07	19	(19.79)
12	0.05	0.01	0.22	0.04	0.49	0.2	19	(18.23)
1	0.42	0.25	0.05	0.19	0.08	0.01	19	(18.18)
22	0.21	0.14	0.1	0.35	0.17	0.02	15	(14.22)
14	0.16	0.15	0.21	0.14	0.17	0.17	15	(9.27)
10	0.19	0.2	0.19	0.16	0.08	0.18	15	(18.29)
7	0.21	0.36	0.14	0.04	0.22	0.03	15	(15.76)
5	0.16	0.26	0.26	0.02	0.01	0.29	15	(15.88)
16	0.04	0.42	0.14	0.19	0.02	0.18	11.5	(10.71)
11	0.19	0.09	0.2	0.21	0.19	0.13	11.5	(10.99)
20	0.41	0.16	0.33	0.03	0.02	0.05	10	(10.90)
4	0.12	0	0.33	0.05	0.26	0.25	9	(9.75)
6	0.06	0.02	0.27	0.23	0.2	0.22	7.5	(6.62)
2	0.14	0.16	0.23	0.2	0.17	0.09	7.5	(8.32)
9	0	0.38	0.13	0	0.3	0.19	6	(5.19)
15	0.13	0.24	0.13	0.12	0.18	0.21	4.5	(8.19)
13	0.27	0.19	0.14	0.16	0.22	0.02	4.5	(3.65)
18	0.36	0.36	0.02	0.09	0.04	0.1	3	(3.80)
17	0.39	0.07	0.14	0.16	0.08	0.14	2	(2.87)
19	0	0.29	0.12	0.06	0.52	0	1	(1.74)

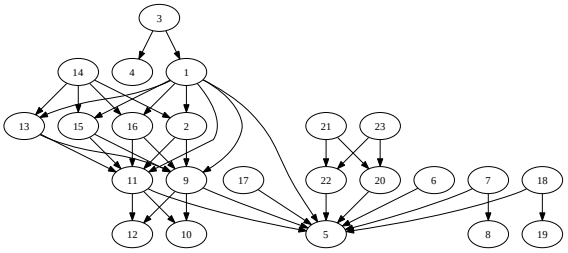
$$\left|\mathcal{V}'^{t}\right| = 23,\ \left|\chi\left(\mathcal{G}'^{t}\right)\right| = 0,\ \kappa\left(\mathcal{G}'^{t}\right) = 1$$

(c) Exemple de *fitness* subjetiu i mesura de consistència per l'aiGA a la iteració $t = 3$.

Figura 3.25: Exemple de *fitness* subjetiu i mesura de consistència per l'aiGA sense cicles a la iteració $t = 3$.



(a) Graf d'ordenació parcial.



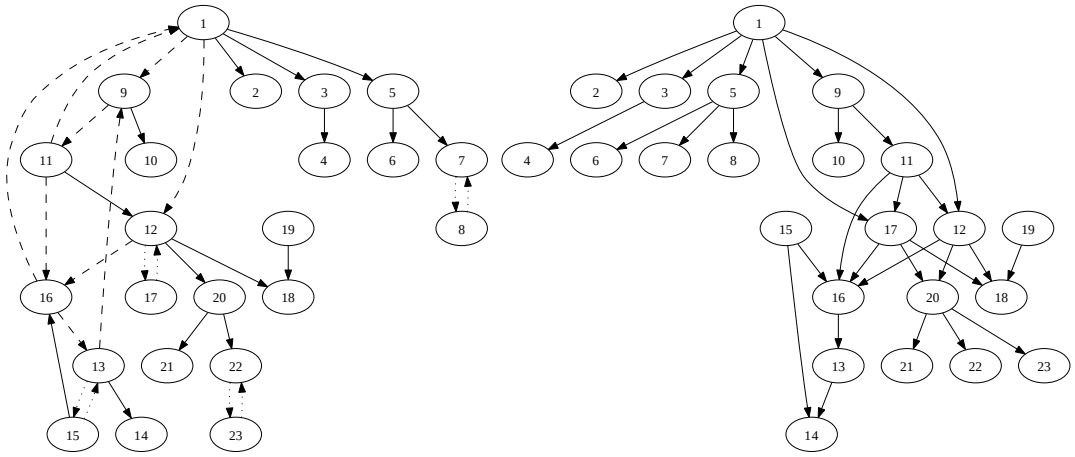
(b) Graf d'ordenació parcial havent eliminat el cicle i normalitzat els empats.

Individu	**DUR.T**	**ENE.C**	**ENE.T**	**MFC.C**	**PIT.C**	**PIT.T**	$\hat{r(v)}$	ε-**SVM** $\hat{r}(v)$
3	0.06	0.13	0.21	0.19	0.29	0.11	23	(4.65)
14	0.21	0.2	0.1	0.14	0.14	0.21	22	(21.10)
1	0.12	0.28	0.13	0.2	0.15	0.12	21	(20.24)
23	0.26	0.06	0.13	0.31	0.12	0.12	19.5	(18.75)
21	0.2	0.15	0.11	0.27	0.05	0.21	19.5	(18.64)
18	0.31	0.06	0	0.01	0.37	0.21	15.5	(14.71)
16	0.06	0.22	0.28	0.25	0.17	0.02	15.5	(14.64)
15	0.21	0.16	0.03	0.2	0.09	0.31	15.5	(16.37)
7	0.26	0.25	0.16	0.02	0.16	0.15	15.5	(16.24)
13	0.35	0.05	0.11	0.07	0.03	0.39	15.5	(14.69)
2	0.01	0.18	0.34	0.16	0.11	0.2	15.5	(14.69)
17	0.2	0.12	0.1	0.14	0.28	0.14	11.5	(12.28)
6	0.08	0.21	0.14	0.21	0.21	0.15	11.5	(12.30)
22	0.13	0.27	0.25	0.12	0.09	0.14	8	(7.26)
20	0.28	0.14	0.29	0.06	0.06	0.16	8	(7.13)
19	0.18	0.27	0.15	0.08	0.23	0.09	8	(8.80)
8	0.12	0.14	0.2	0.19	0.2	0.16	8	(7.16)
4	0.01	0.17	0.25	0.24	0.26	0.07	8	(11.02)
9	0.21	0.21	0.27	0.07	0.13	0.12	4.5	(4.42)
11	0.14	0.17	0.19	0.13	0.24	0.13	4.5	(5.33)
12	0.22	0.1	0.25	0.13	0.11	0.19	2.5	(5.50)
10	0.11	0.26	0.26	0.22	0.1	0.05	2.5	(14.27)
5	0.15	0.38	0.15	0.08	0	0.24	1	(1.79)

$$\left|\mathcal{V}'^t\right| = 23, \left|\chi\left(\mathcal{G}'^t\right)\right| = 3, \kappa\left(\mathcal{G}'^t\right) = 0.87$$

(c) Exemple de *fitness* subjetiu i mesura de consistència per l'aiGA a la iteració $t = 3$.

Figura 3.26: Exemple de *fitness* subjetiu i mesura de consistència per l'aiGA amb un cicle a la iteració $t = 3$.



(a) Graf d'ordenació parcial.

(b) Graf d'ordenació parcial havent eliminat els cicles i normalitzat els empats.

Individu	**DUR.T**	**ENE.C**	**ENE.T**	**MFC.C**	**PIT.C**	**PIT.T**	$\hat{r(v)}$	ε-**SVM** $\hat{r}(v)$
1	0.11	0.24	0.05	0.01	0.24	0.36	23	(22.17)
9	0.11	0.07	0.01	0.03	0.5	0.29	22	(21.23)
11	0.01	0.3	0.03	0.09	0.3	0.27	21	(20.24)
17	0.32	0.28	0.06	0.05	0.19	0.07	19.5	(18.59)
12	0.18	0.12	0.25	0.21	0.04	0.19	19.5	(18.76)
15	0.04	0.16	0.21	0.2	0.11	0.27	18	(18.89)
5	0.08	0.31	0.12	0	0.29	0.19	17	(17.71)
19	0.12	0.35	0.09	0.1	0.26	0.08	16	(16.85)
3	0.11	0.21	0.2	0.14	0.11	0.22	15	(14.20)
2	0.2	0.17	0.23	0.19	0.21	0	14	(13.19)
20	0.52	0.2	0.05	0.09	0.05	0.09	10.5	(11.35)
10	0.12	0.17	0.21	0.1	0.24	0.17	10.5	(9.73)
8	0.08	0.25	0.04	0.05	0.39	0.18	10.5	(10.21)
7	0.22	0.04	0.09	0.1	0.24	0.31	10.5	(9.62)
6	0.04	0.08	0.23	0.21	0.26	0.18	10.5	(9.71)
4	0.16	0.34	0.03	0.26	0.05	0.17	10.5	(11.30)
16	0.23	0.07	0.1	0.2	0.2	0.2	7	(6.26)
23	0.27	0	0.47	0	0.12	0.13	4	(3.28)
22	0.27	0.29	0.31	0.05	0.06	0.02	4	(3.24)
21	0.25	0.08	0.42	0.12	0.08	0.05	4	(4.82)
13	0.2	0.2	0.23	0.09	0.17	0.11	4	(9.10)
18	0.17	0.17	0.03	0.01	0.22	0.38	4	(4.87)
14	0.12	0.06	0.09	0.16	0.32	0.25	1	(5.84)

$$\left|\mathcal{V}'^t\right| = 23,\ \left|\chi\left(\mathcal{G}'^t\right)\right| = 6,\ \kappa\left(\mathcal{G}'^t\right) = 0.74$$

(c) Exemple de *fitness* subjetiu y mesura de consistència per l'aiGA a la iteració $t = 3$.

Figura 3.27: Exemple de *fitness* subjetiu i mesura de consistència per l'aiGA amb un multicicle a la iteració $t = 3$.

robusta. A tal efecte es va analitzar l'ajust de l'iGA simple (Alías *et al.*, 2003, 2004) per tal d'establir un marc de treball comú. Posteriorment, es va realitzar el mateix anàlisi en unes noves proves d'ajust que seguien el mateix disseny que les proves anteriors i substituien l'iGA simple per un iGA actiu (aiGA) adaptat al problema de l'ajust de pesos mitjançant CTP-SU (Alías *et al.*, 2006*a*).

Millora de la consistència dels usuaris avaluadors

Per tal d'obtenir com a punt de partida, uns indicadors de la consistència de l'ajust amb iGA es va calcular la mètrica $\kappa\left(\mathcal{G}'^{t_f}, w\right)$ (apartat 3.4.7 - on t_f indica el temps final del procés iteratiu d'ajust dels pesos de selecció) a partir del registre de les proves anteriors (Alías *et al.*, 2003).

En aquest apartat es recull l'experimentació realitzada a (Llorà *et al.*, 2005*a*; Alías *et al.*, 2006*a*). Els resultats de les mètriques obtingudes es detallen a la part superior de la taula 3.3. Com a primera apreciació es pot observar el fet que només un usuari *expert* va ser consistent en una sola frase (prova). Addicionalment, es va analitzar el comportament temporal de la mètrica $\kappa\left(\mathcal{G}'^{t_i}, w\right)$ a través de les i generacions tal com es mostra les figures 3.28(a), 3.28(c), 3.28(e) i 3.28(g). Segons aquestes figures es pot observar com tots els usuaris, independentment del seu perfil - *novell*, *especialista* o *expert*- van tenir problemes per a mantenir un criteri consistent al llarg de l'experiment usant l'iGA simple. Això és degut, en gran mesura, al gran nombre d'avaluacions necessàries abans d'aconseguir la convergència de l'iGA. Un altre descobriment important d'aquesta anàlisi fou observar la presència d'inconsistències ja des de l'inici del procés iteratiu de l'iGA. En mitjana, els usuaris van tendir a contradir-se al voltant del torneig 14, amb un valor mitjà de 2.83 de contradiccions per execució. Aquestes inconsistències provoquen un entorn excessivament sorollós en l'optimització de la funció de *fitness*, provocant un increment del número d'avaluacions necessàries per tal d'obtenir solucions d'elevada qualitat (Miller i Goldberg, 1995; Goldberg, 2002; Sastry i Goldberg, 2002). Això provoca l'augment de la fatiga de l'usuari i fa que el procés d'aprenentatge perdi efectivitat.

Per tal de validar l'eficiència teòrica de l'aiGA, es va repetir l'experiment realitzat amb anterioritat (Alías *et al.*, 2003), però aquesta vegada reemplaçant l'iGA simple del procés per l'iGA actiu (aiGA) (Llorà *et al.*, 2005*b*) segons l'esquema de la figura 3.20 de l'apartat 3.4.6. La part inferior de la taula 3.3 mostra els resultats del càlcul de la mesura de consistència $\kappa\left(\mathcal{G}'^{t_f}, w\right)$ per als tres perfils d'usuari. Així mateix, les figures 3.28(b), 3.28(d), 3.28(f) i 3.28(h) mostren una evolució temporal de $\kappa\left(\mathcal{G}'^{t}, w\right)$ al llarg dels tornejos. A simple vista, es pot observar com reemplaçar l'algorisme genètic interactiu simple per l'interac-

tiu actiu permet augmentar decididament la consistència de les avaluacions dels usuaris, i ajuda a mantenir el seu criteri d'avaluació, ja que tan sols dos dels dotze experiments finalitzaren inconsistentment, és a dir, $\kappa(\mathcal{G}'^{t_f}, w) < 1$ (veure la part inferior de la taula 3.3). No obstant això, la consistència de les execucions usant aiGA està molt per sobre de l'aconseguida usant l'iGA simple, on només un usuari (l'*expert*) va ser capaç de ser consistent al llarg de tota la prova. Una altra de les observacions interessants que s'extreuen de l'anàlisi dels resultats usant iGA actiu és que, fins i tot quan l'usuari comet alguna contradicció en les avaluacions, la selecció *activa* dels tornejos de candidats basat en el graf d'ordenació parcial $\mathcal{G}'$ ajuda a l'usuari a tornar al camí de la consistència (veure com a exemple la figura 3.28(h), on l'usuari novell recupera la consistència al voltant del torneig 45).

A partir d'aquests resultats, es pot calcular la millora en la consistència aconseguida al reemplaçar l'iGA simple per l'iGA actiu en el procés d'ajust de pesos de la funció de cost mitjançant la interfície interactiva (veure taula 3.4). Com a conclusió, es pot observar que gràcies a usar aiGA la consistència al llarg del procés evolutiu millora de forma evident, ajudant a l'usuari perquè pugui avaluar les solucions proposades de forma consistent.

iGA simple **Frase**	**Usuari Novell**	**Usuari Especialista**	**Usuari Expert**
"De la seva selva"	0.944	0.855	0.784
"Fusta de Birmània"	0.857	0.769	0.911
"I els han venut"	0.894	0.867	0.731
"Grans extensions"	0.942	0.800	1.000
iGA actiu **Frase**	**Usuari Novell**	**Usuari Especialista**	**Usuari Expert**
"De la seva selva"	1.000	0.892	1.000
"Fusta de Birmània"	1.000	1.000	1.000
"I els han venut"	1.000	1.000	0.948
"Grans extensions"	1.000	1.000	1.000

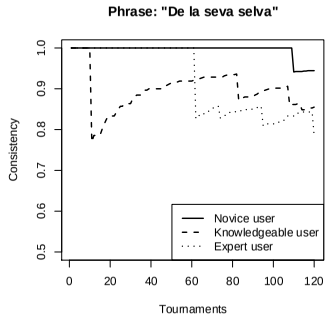
Taula 3.3: Consistència final $\kappa\left(\mathcal{G}'^{t_f}, w\right)$ (equació (3.5)), segons el perfil d'usuari, per a les quatre frases de l'experiment (Llorà *et al.*, 2005*a*; Alías *et al.*, 2006*a*).

iGA simple **Frase**	**Usuari Novell**	**Usuari Especialista**	**Usuari Expert**	**Usuari Valor mitjà**
"De la seva selva"	5.89%	4.30%	27.50%	12.56%
"Fusta de Birmània"	16.67%	30.01%	9.81%	18.83%
"I els han venut"	11.91%	15.00%	29.76%	18.89 %
"Grans extensions"	6.12%	25.00%	0,00%	10.37%
Valor mitjà	10.15%	18.58%	16.77%	15.16 %

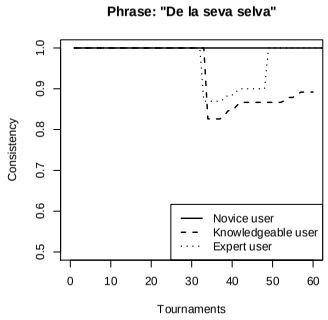
Taula 3.4: Augment de la consistència aconseguida quan es reemplaça l'iGA simple per l'iGA *actiu*, calculat com la diferència absoluta entre les consistències de cada mètode presentades a la taula 3.3(Llorà *et al.*, 2005*a*; Alías *et al.*, 2006*a*).

Frase	**Usuari Novell**	**Usuari Especialista**	**Usuari Expert**	**Usuari Valor mitjà**
"De la seva selva"	2.00	2.00	2.00	2.00
"Fusta de Birmània"	2.00	2.00	2.00	2.00
"I els han venut"	2.00	2.00	2.00	2.00
"Grans extensions"	2.00	2.67	2.00	2.23
Valor mitjà	2.00	2.17	2.00	2.06

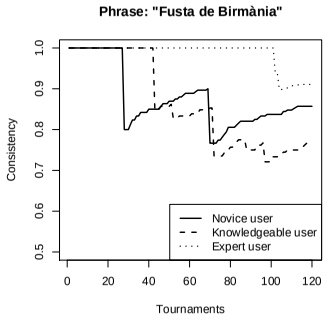
Taula 3.5: Millora de l'eficiència aconseguida quan es reemplaça l'iGA simple per l'iGA *actiu*, calculada com el quocient entre el número de tornejos necessaris abans de convergir(Llorà *et al.*, 2005*a*; Alías *et al.*, 2006*a*).



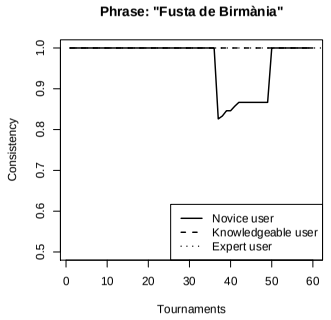
(a) Algorisme interactiu simple (iGA).



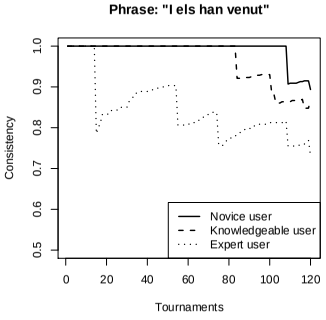
(b) Algorisme interactiu actiu (aiGA).



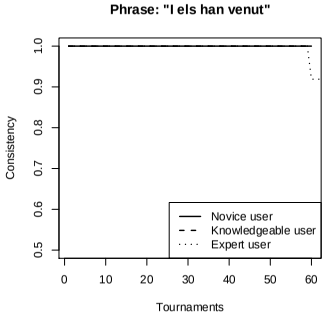
(c) Algorisme interactiu simple (iGA).



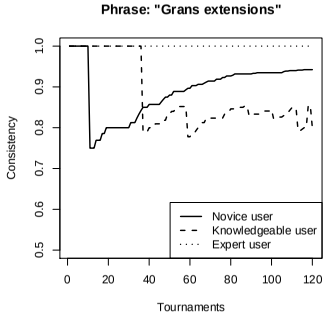
(d) Algorisme interactiu actiu (aiGA).



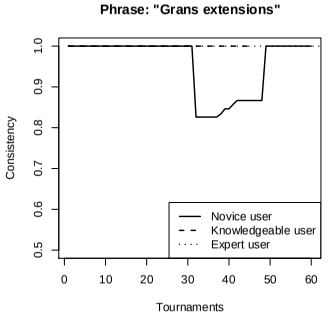
(e) Algorisme interactiu simple (iGA).



(f) Algorisme interactiu actiu (aiGA).



(g) Algorisme interactiu simple (iGA).



(h) Algorisme interactiu actiu (aiGA).

Figura 3.28: Evolució de la consistència d'usuari avaluada mitjançant la mesura $\kappa\left(\mathcal{G}'^{t}, w\right)$ per a les locucions *"de la seva selva"*, *"fusta de birmània"*, *"i els han venut"* i *"grans extensions"*. Les figures comparen l'evolució de la consistència per a diferents perfils d'usuari usant l'algorisme interactiu simple (iGA) o l'algorisme interactiu actiu (aiGA)(Llorà *et al.*, 2005*a*; Alías *et al.*, 2006*a*).

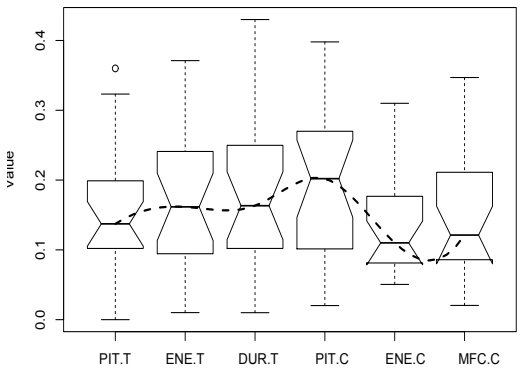
Millora de l'eficiència de l'entrenament

Una de les raons fonamentals del disseny dels aiGA es centra en aconseguir una reducció important del nombre d'avaluacions que ha de realitzar l'usuari, de forma que s'aconsegueixi una reducció de la seva fatiga (Llorà *et al.*, 2005*b*). Tot i que l'objectiu fonamental de la aplicació dels aiGA al problema d'ajust subjectiu dels pesos de la funció de cost fou augmentar la consistència de les avaluacions de l'usuari, substituir els iGA simples per l'iGA actiu permet, a la vegada, reduir el nombre d'iteracions del procés evolutiu (Alías *et al.*, 2006*a*). El criteri de parada, tal com s'explica a (Alías *et al.*, 2003), el marca l'usuari quan és incapaç d'escoltar diferències entre les solucions que se li presenten, cosa que provoca que el nombre d'empats augmenti de manera considerable. Encara que no es va prendre cap mesura especial per a millorar l'eficiència del procés, els resultats obtinguts són realment bons. Concretament, la taula 3.5 mostra una reducció mitjana del 50% en el nombre d'avaluacions que ha de realitzar un usuari abans de convergir, de tal manera que les proves de l'aiGA van durar, en valor mitjà, la meitat que les realitzades anteriorment mitjançant l'iGA simple; es va passar d'unes 6 generacions - uns 120 tornejos - a unes 3 generacions - uns 60 tornejos -. Aquesta millora es pot apreciar a la figura 3.28. Així doncs, incorporar l'aiGA com a algorisme de base en el procés d'entrenament dels pesos permet atacar satisfactoriament dos dels problemes fonamentals en qualsevol procés d'ajust interactiu: la consistència i la fatiga de l'usuari; qüestions que, com es veurà en l'anàlisi següent, permeten a la vegada, aconseguir unes solucions de millor qualitat.

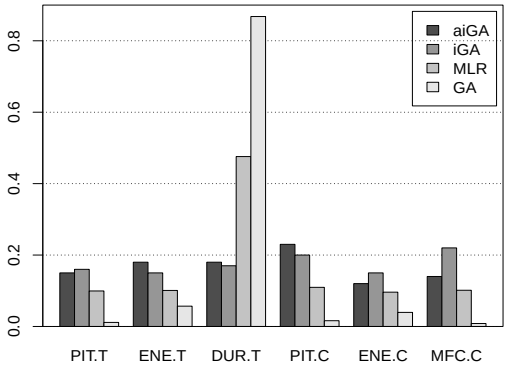
Pesos obtinguts i diferència de comportament respecte els altres mètodes

Un cop validada la capacitat de l'aiGA en obtenir un ajust de pesos de manera robusta, eficient i fiable en sistemes CTP-SU, es van analitzar i comparar els patrons de pesos obtinguts mitjançant aiGA amb els obtinguts mitjançant les altres metodologies considerades, fossin obtinguts de manera perceptiva (iGA) o de manera automàtica amb distàncies cepstrals (GA, MLR).

Els pesos obtinguts dels diferents usuaris per a les diferents frases mitjançant aiGA es mostren a la figura 3.29(a). En l'esmentada figura es pot observar com per primer cop, a diferència dels altres mètodes, el pes més important no és w_3 =DUR.T sinó que passa a ser la diferència de *pitch* en el punt de concatenació w_4 =PIT.C. De tota manera, si s'analitzen les diferències entre els diferents pesos en termes de significança (3.29(c)) es pot observar que cap combinació de parelles de pesos supera la prova de significança (*t*-Student per parelles), per tant no es pot afirmar que els pesos siguin clarament diferents entre sí. En



(a) *Boxplots* dels valors dels pesos al finalitzar les 3 generacions de l'aiGA.

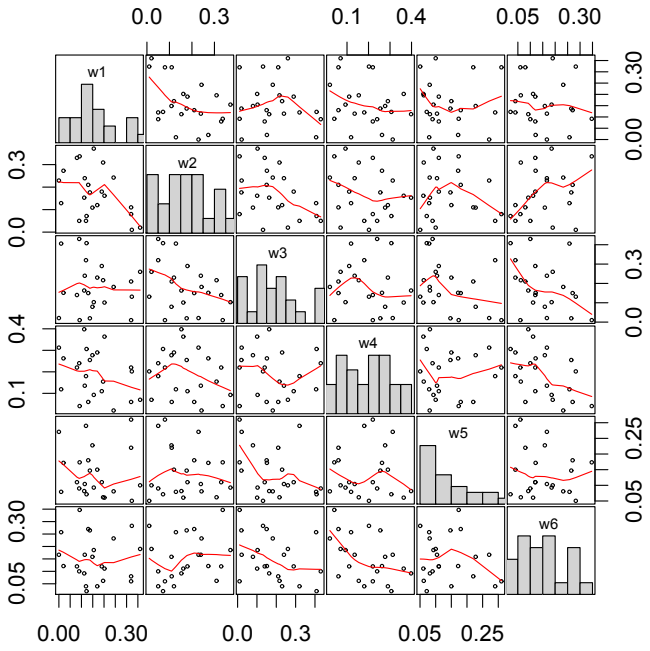


(b) Comparació dels pesos obtinguts en aplicar aiGA amb els pesos obtinguts mitjançant iGA i els processos automàtics MLR i GA.

	PIT.T	ENE.T	DUR.T	PIT.C	ENE.C	MFC.C
PIT.T	1	0.5674	0.5577	0.5652	0.7025	0.9675
ENE.T	0.5674	1	0.9631	**0.9903**	0.3104	0.5654
DUR.T	0.5577	0.9631	1	0.9531	0.3198	0.5568
PIT.C	0.5652	**0.9903**	0.9531	1	**0.3001**	0.5621
ENE.C	0.7025	0.3104	0.3198	**0.3001**	1	0.6391
MFC.C	0.9675	0.5654	0.5568	0.5621	0.6391	1

(c) Significança de les diferències (*t*-Student per parelles) entre els diferents pesos obtinguts.

Figura 3.29: Diferents resultats obtinguts amb aiGA durant el transcurs de 3 generacions d'ajust interactiu juntament amb la seva significança de diferències (*t*-Student per parelles) entre els diferents pesos obtinguts.



$$R_w = \begin{pmatrix} 1.00 & \mathbf{-0.42} & -0.11 & \mathit{-0.30} & -0.05 & -0.07 \\ \mathbf{-0.42} & 1.00 & \mathbf{-0.41} & \mathit{-0.23} & -0.12 & \mathit{0.25} \\ -0.11 & \mathbf{-0.41} & 1.00 & -0.10 & \mathit{-0.38} & \mathit{-0.36} \\ \mathit{-0.30} & \mathit{-0.23} & -0.10 & 1.00 & 0.02 & \mathbf{-0.47} \\ -0.05 & -0.12 & \mathit{-0.38} & 0.02 & 1.00 & -0.16 \\ -0.07 & \mathit{0.25} & \mathit{-0.36} & \mathbf{-0.47} & -0.16 & 1.00 \end{pmatrix}$$

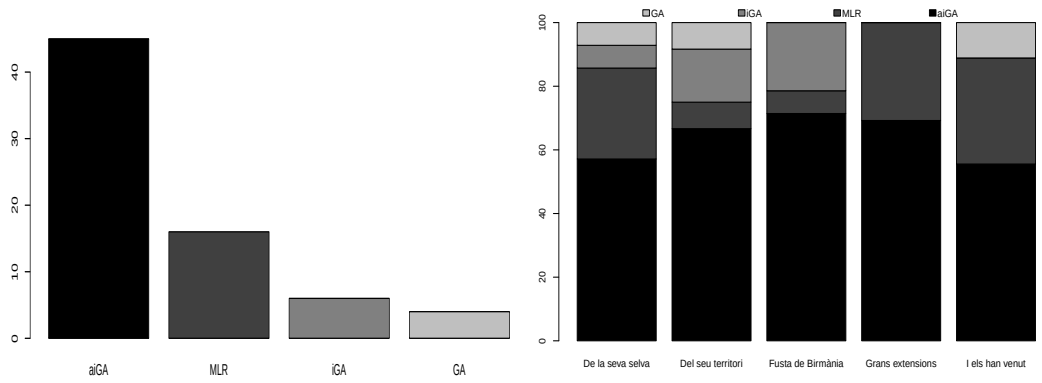
Figura 3.30: Correlacions lineals entre els valors dels pesos després de finalitzar les 4 generacions de l'aiGA, on w_1 =PIT.T, w_2 =ENE.T, w_3 =DUR.T, w_4 =PIT.C, w_5 =ENE.C, w_6 =MFC.C.

l'esmentada taula el valor màxim i el valor mínim s'han ressaltat amb negreta.

Com a última apreciació, es pot observar com el patró de pesos obtingut amb aiGA segueix el mateix comportament que els pesos obtinguts amb iGA pel que fa a la suavització dels valors de pesos (veure figura 3.29(b)). Aquest fet contrasta amb l'aproximació excessivament discriminatòria dels mètodes automàtics que donen una importància molt elevada al valor de pes de la durada de *target*.

Per tal d'analitzar si el comportament dels pesos entre sí és lineal, s'ha analitzat l'impacte de la variació del valor d'un pes respecte als altres pesos a part de veure si aquesta correlació és lineal (figura 3.30). En aquest cas, es pot observar que cap correlació és plenament lineal ni en termes de correlació directa (més gran que zero) o correlació inversa (més petita que zero). En general, es poden apreciar tres lleugeres aproximacions a la correlació inversa en les parelles de pesos següents: <PIT.T,ENE.T>=-0.42,<ENE.T,DUR.T>=-0.41 i <PIT.C,MFC.C>=-0.47 que pot donar a entendre que aquests parelles de pesos competeixen entre elles (en termes de supervivència quan un pes incrementa el seu valor l'altra es veu forçat a decrementar-lo). En general es pot afirmar que cap dels comportaments dels pesos entre sí són lineals.

Comparació perceptiva dels diferents mètodes d'ajust de pesos (MLR / GA / iGA / aiGA)



(a) Vots entre els diferents mètodes.

(b) Comparativa dels diferents mètodes detallats per frase.

Figura 3.31: Comparativa perceptiva entre els diferents mètodes d'ajust de pesos pel corpus *url_fer_ct*.

Per últim, es va voler realitzar una validació perceptiva dels diferents pesos obtinguts

mitjançant les diferents metodologies d'ajust considerades (aiGA, iGA, GA i MLR) i a tal efecte es van considerar les mateixes frases que s'havien emprat per tal de realitzar l'ajust i es va sintetitzar cadascuna d'elles segons els diferents conjunts de pesos obtinguts amb cadascuna de les metodologies d'ajust. Seguidament, es va demanar a un grup de 10 usuaris que avaluessin les diferents síntesis per parelles i escollissin quina era la síntesi en la que apreciaven una millor naturalitat. Els resultats de les votacions s'exemplifiquen en la figura 3.31, on es poden observar els resultats de la puntuació dels mètodes globals (figura 3.31(a)) i detallats per frase (figura 3.31(b)).

Analitzant els resultats de les avaluacions globals es pot apreciar que aiGA és l'única metodologia d'ajust evolutiva capaç de millorar la qualitat sintètica del mètode d'ajust no evolutiu MLR (en aquesta tesi, considerat el *baseline* dels processos d'ajust de pesos en sistemes CTP-SU). Es pot observar també que la millora de l'aiGA respecte els altres mètodes és significativa amb uns valors de $p = 6.93 \cdot 10^{-11}$ respecte l'MLR, $p < 10^{-12}$ respecte l'iGA i $p < 10^{-12}$ respecte el GA automàtic. També es pot observar que malgrat que els pesos d'iGA i aiGA poden resultar semblants, petites diferències en l'ajust de pesos (veure figura 3.29(b)) poden comportar grans diferències en la qualitat perceptiva de la síntesi realitzada.

3.5 Aspectes de millora en l'ajust dels pesos mitjançant aiGA

En aquest capítol s'han explicat les diferents aproximacions d'optimització evolutiva que s'han emprat per ajustar els pesos de la funció de cost del mòdul de selecció en sistemes CTP-SU. Primerament, s'ha argumentat la idoneïtat d'aplicar l'optimització evolutiva en la cerca dels valors apropiats de pesos de la funció de cost de selecció i com aquesta permetia *i*) trencar el comportament lineal típic de l'ajust per mínims quadrats (MLR) i *ii*) incorporar la cooperació de la percepció humana com un actor essencial en determinar la qualitat de les síntesis realitzades. Un cop validat el principi de viabilitat (*proof-of-principle*) d'emprar aiGAs per realitzar l'ajust de pesos, queden qüestions per resoldre en l'ajust de pesos que necessiten d'una recerca més exhaustiva d'un corpus de mida més gran de l'emprada fins al moment.

Bàsicament la recerca que s'ha de realitzar en l'aplicació de l'aiGA per l'ajust de pesos mitjançant selecció d'unitats ha de girar segons els eixos que s'exposen a continuació:

i) No totes les unitats tenen perquè seguir el mateix patró de pesos. Les unitats diferents que composen un corpus de veu tenen especifitats diferents a nivell fonètic (Campillo *et al.*, 2005): el seu punt d'articulació, mode d'articulació, així com la seva sonoritat o

tipologia de vocal poden determinar que els pesos que són bons per un grup (clúster) d'unitats no siguin bons per altres grups d'unitats. És per aquest motiu que cal un estudi amb més profunditat de quins pesos s'han d'aplicar a quins grups d'unitats per tal de determinar a quin nivell (global, unitat, etc.) s'ha de fer l'ajust. En aquesta qüestió s'ha de tenir en compte que no tots els mètodes d'ajust permeten arribar a la màxima profunditat de l'ajust: mentre que els mètodes automàtics permeten realitzar l'ajust emprant tanta computació com sigui necessària, els mètodes d'ajust interactius (perceptius) adopten la restricció de la fatiga humana, i com a conseqüència no permet realitzar l'ajust de pesos a baix nivell.

ii) No tots els usuaris obtenen el mateix patró de pesos per un disseny concret d'una prova (diferents criteris). Cal doncs, estudiar la manera de fusionar els ajustos dels diferents usuaris per a poder obtenir un ajust de pesos consensuat entre els usuaris per tal de garantir una màxima qualitat perceptiva per tots ells.

iii) L'índex de consistència permet descartar ajustos que hagin resultat contradictoris, poc robustos, o sorollosos a l'hora d'obtenir el patró de pesos final. De tota manera, resulta necessari obtenir indicadors més enllà de la consistència que determinin la qualitat d'un ajust, el moment de finalitzar una prova (ha convergit o ha resultat encallada). A tal efecte es poden tenir en compte indicadors tals com el nombre d'empats, per mesurar l'ambigüitat, o la semblança dels millors pesos entre ells, per analitzar la multimodalitat del problema.

iv) En les proves descrites fins ara s'ha provat la viabilitat d'ajustar 6 pesos per ponderar diferents subcostos de tipologia acústica en un corpus de mida molt limitada (8 minuts). En aquest sentit, no es disposa d'informació per analitzar el comportament de l'aiGA en un entorn real de selecció d'unitats que ajusti diferents tipologies de subcostos (barrejar subcostos lingüístics amb subcostos acustics) i amb molta variabiliat d'elecció dins del corpus ($>$ 1h). Per això resulta necessari analitzar la viabilitat la metodologia de l'aiGA en un entorn real de selecció d'unitats que contempli aquests aspectes.

En la segona part d'aquesta tesi doctoral, s'analitzen les diferents contribucions que s'han fet per tal de solventar les diferents qüestions que s'acaben de plantejar i així realitzar un estudi amb completitud de com s'ha d'emprar l'aiGA en un entorn real de selecció d'unitats de manera robusta i eficient. En aquest sentit es vol realitzar el pas que manca per passar d'una prova de viabilitat (Alías *et al.*, 2006*a*; Alías, 2006) a una aplicabilitat general de l'ajust de pesos mitjançant aiGA.

Part II

Contribucions a l'ajust de pesos per sistemes CTP-SU

CAPÍTOL 4

Prova de viabilitat: corpus petit amb subcostos acústics

4.1 Introducció

En aquest capítol es presenta un estudi amb detall de diferents problemàtiques que presenten les metodologies explicades d'ajust de pesos independentment siguin aquestes de tipologia automàtica o perceptiva. Primer, es presenta una descripció detallada del corpus de veu utilitzat: *i*) la seva composició, *ii*) la seva variabilitat prosòdica, i *iii*) la distribució dels seus subcostos juntament amb la seva normalització i transformació. Després es procedeix a estudiar en profunditat l'ajust de pesos automàtic de l'esmentat corpus, explicat en l'apartat 2.4.2, focalitzant l'estudi en els problemes de consistència i fiabilitat que presenta l'optimització dels pesos en un entorn sorollós. En aquest sentit, es discuteix la problemàtica d'ajustar els pesos de manera global sense respectar les seves especificitats fonètiques. Aquesta problemàtica (nivell d'ajust) ve donada pel fet que els subcostos cobren diferent importància en funció de la unitat que es vulgui recuperar (Campillo *et al.*, 2005). En aquest sentit, es presenta una segona contribució proposant un nivell d'ajust intermedi per grups d'unitats basada en arbres de decisió (CART) i fent servir certs índex de bondat per determinar el nombre de clústers (grups) triat. A posteriori, previ a l'ajust perceptiu pròpiament dit, es presenten dues variants de la metodologia perceptiva a efectes de millorar la seva robustesa: *i*) les modificacions de l'algorisme de generació de la forma d'ona, per tal de poder avaluar d'una manera fidel el funcionament de la selecció

d'unitats, i *ii*) la definició de nous indicadors del procés evolutiu per tal de tenir informació de l'ambigüitat en l'execució, la convergència d'un usuari i la correlació entre diferents usuaris, més enllà de la mesura de consistència descrita a l'apartat 3.4.7.

Posteriorment, es duen a terme els ajustos perceptius basats en l'optimització dels pesos per cadascun del clústers definits. Els resultats obtinguts són validats respecte les tècniques d'ajust automàtic mitjançant un test perceptiu que es presenta a un altre grup d'usuaris diferent del que ha realitzat l'ajust a través d'una puntuació d'opinió mitjana comparativa (*Comparison Mean Opinion Score* - CMOS).

A continuació, els resultats del CMOS s'utilitzen per obtenir uns altres pesos mitjançant una de les tècniques d'ajust perceptiu vigents, el MOS-*Postmapping* (Chu i Peng, 2001; Peng *et al.*, 2002) (explicada a l'apartat 2.4.3). Un cop obtinguts els nous pesos, es torna a repetir l'avaluació comparativa considerant ambdós mètodes perceptius (aiGA i MOS-*Postmapping*) i es presenten els resultats obtinguts, juntament amb les seves conclusions i les noves línies d'investigació que apareixen a partir d'aquestes contribucions al problema de l'ajust de pesos.

4.2 Descripció del corpus

En d'aquest capítol, s'utilitza el mateix corpus emprat en la recerca explicada fins ara (apartats 3.2.4, 3.3.3 i 3.4.8). El corpus, referenciat com a *bdp2* (Guaus i Iriondo, 2000*a*,*b*) - posteriorment etiquetat com *url_fer_ct* - es va dissenyar per formar part d'un sintetitzador per difonemes (2a generació) en català. El corpus està composat de 1207 unitats, de les quals 895 són difonemes i 312 són trifonemes. Aquest conjunt inicial es va estendre amb 313 frases (de 8 min. de durada) balancejades fonèticament per a obtenir més variabilitat d'unitats, i així disposar d'una primera aproximació de corpus on realitzar una selecció d'unitats.

Per tal d'optimitzar la selecció d'unitats, el primer pas és estudiar la tipologia de dades que determinen el procés de selecció. A tal efecte, en aquest apartat es descriu el corpus emprat en les primeres proves d'aquesta tesi doctoral. La descripció inclou *i*) la composició del corpus emprat tant a nivell d'al·lòfons com d'unitats (difonemes i trifonemes), *ii*) les funcions de distribució dels diferents paràmetres prosòdics, considerant també la seva adequació a la distribució normal, i *iii*) les distribucions dels diferents subcostos emprats, considerant també la seva normalització (entre 0 i 1) amb l'objectiu d'integrar-los en la funció de cost detallada en l'apartat 2.2.3.

4.2.1 Composició fonètica i prosòdica

El corpus *url_fer_ct* es composa per 37 al·lòfons diferents que es detallen en la figura 4.1, els quals es poden dividir en 28 al·lòfons sonors i 8 al·lòfons sords. Alhora, els 28 al·lòfons sonors es divideixen en 18 al·lòfons consonàntics, 8 de vocàlics i 2 semivocàlics. Per a més informació, es pot consultar a l'annex C la taula amb la distribució de fonemes (taula C.2) així com la taula de tipificació fonètica (tipologia, punt d'articulació, mode d'articulació i sonoritat) dels al·lòfons (taula C.1). Segons aquestes dades, l'al·lòfon amb major presència en el corpus (així com en llengua catalana Rafel (1980)) és la vocal neutre /@/ amb un 15.38% de presència dins el corpus. Seguidament s'hi troba la /i/ amb un 7.2%, i les consonants /s/,/n/,/l/ i /t/ amb un 5.97%, 5.74%, 5.33% i 4.68% respectivament.

La mateixa anàlisi es pot realitzar per la cobertura d'unitats (difonemes i trifonemes) en el corpus, en aquest cas el nombre d'unitats creix fins a 1207 unitats diferents, fent un total de 9863 unitats enregistrades. A la figura 4.2 es pot observar la distribució de les diferents unitats en el corpus considerat. En aquest cas també es pot observar la predominança de la vocal neutra en les unitats més representades en el corpus: concretament en 16 de les 20 unitats amb més cobertura (les excepcions són /si/,/ns/,/un/ i /st/). En l'annex C es pot trobar la taula detallada de cobertura fonètica (taula C.3).

	Max	Min	Mitjana	Mediana	Desviació típica
pitch (Hz)	238	65	117.44	115	16.30
Δ *pitch* (Hz)	110	-130	-0.64	-2	15.62
energia (RMS)	0.01	0	$3.5{\cdot}10^{-3}$	$3.7{\cdot}10^{-3}$	$2.1{\cdot}10^{-3}$
Δenergia (RMS)	0.01	-0.01	$-1{\cdot}10^{-4}$	$-3{\cdot}10^{-4}$	$2.9{\cdot}10^{-3}$
ritme (*z-scored*)	10.18	-4.05	0	-0.18	1
Δ ritme (*z-scored*)	6.28	-8.84	0.07	0.05	1.13

Taula 4.1: Estadístiques de primer ordre del corpus *url_fer_ct*.

A nivell prosòdic, es realitza un estudi de la naturalesa de les dades per conèixer-ne la noció estadística. Aquest estudi permet disposar d'un coneixement més acurat per optimitzar el càlcul de la funció de cost, que en aquest capítol es compon íntegrament per subcostos acústics. Els subcostos acústics, referenciats a l'apartat 2.3.4 com a ASF, es diferencien dels de naturalesa simbòlica (IFF) perquè es deriven d'una parametrització prosòdica prèvia del corpus.

Per a l'anàlisi es consideren els tres paràmetres bàsics (Campbell i Black, 1997) que són

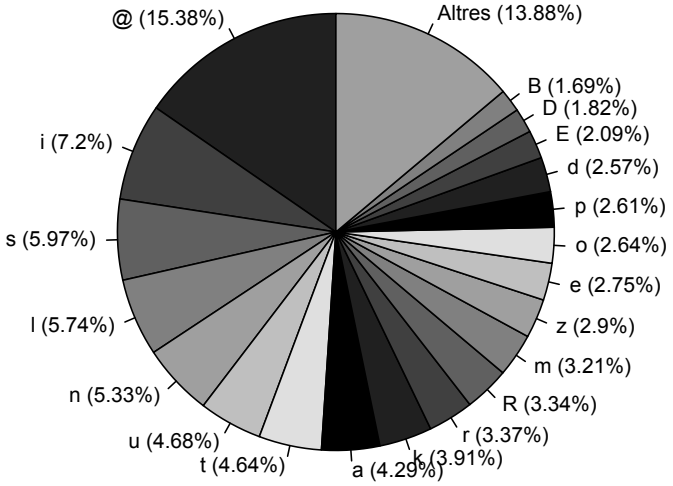


Figura 4.1: Distribució dels diferents fonemes en notació SAMPA (Wells *et al.*, 1992) en català a dins del corpus *url_fer_ct*.

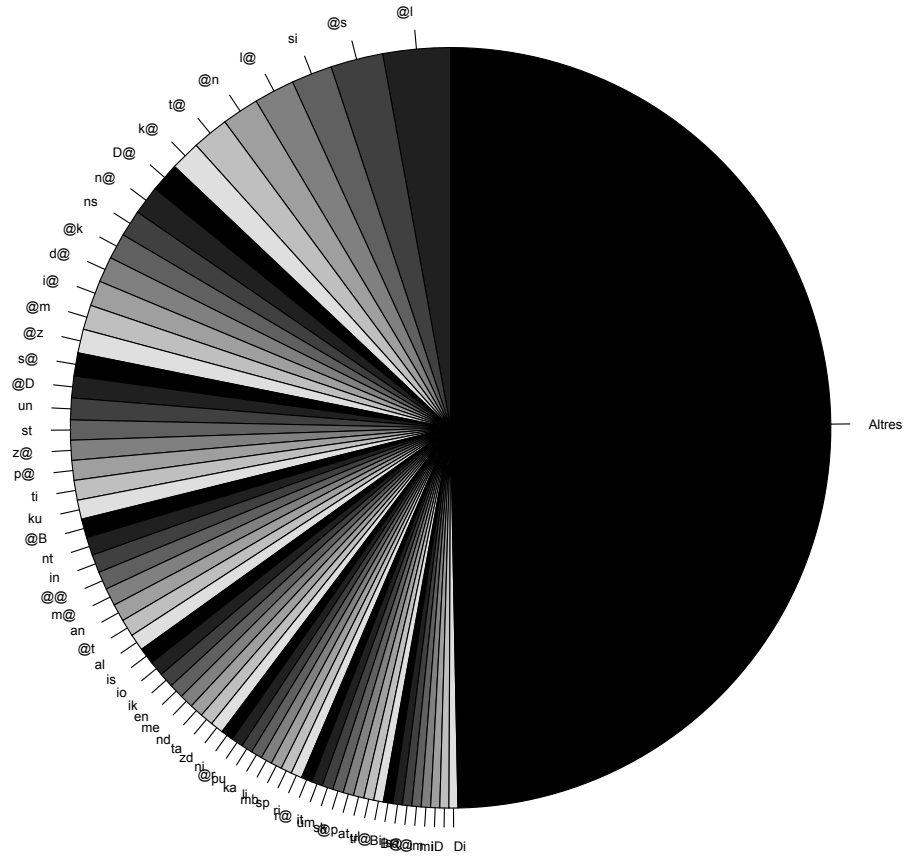


Figura 4.2: Distribució de les diferents unitats en notació SAMPA (Wells *et al.*, 1992) en català pel corpus *url_fer_ct*.

	Asimetria (*skewness*)	Curtosis	Test de Lilliefors
pitch	0.77	1.6535	0.1055
Δ *pitch*	0.201	3.1172	0.0726
energia	0.2309	-0.5436	0.0753
Δenergia	0.2152	-0.5478	0.0424
ritme	1.3235	4.0822	0.0863
Δ ritme	0.1268	2.8315	0.0433

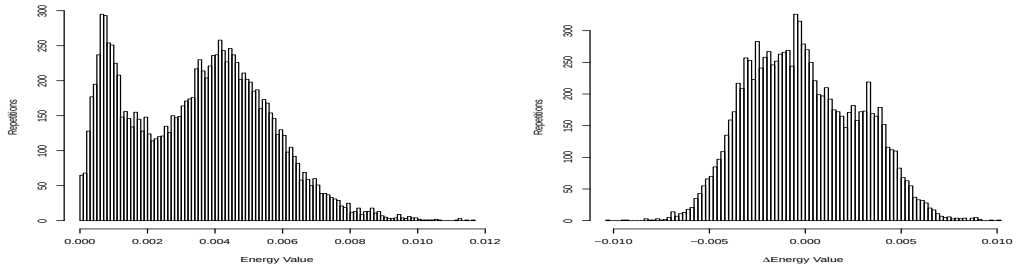
Taula 4.2: Estadístiques de segon ordre i proves de normalitat del corpus *url_fer_ct*.

(veure apartat 2.2.2): *i*) el *pitch*, *ii*) l'energia i *iii*) la durada, així com les seves derivades. La derivada es calcula com la variació del paràmetre prosòdic entre dues unitats consecutives. El *pitch* i l'energia es calculen directament a partir del senyal del veu, essent les seves unitats: Hz (Hertzs) pel *pitch* i RMS per l'energia (Hirschberg i Nakatani, 1998). El ritme de la parla es mesura mitjançant el càlcul de la durada dels al·lòfons filtrada segons una normalització *z-score* (Navas *et al.*, 2002*b*) degut a l'alta dependència del seu valor respecte la identitat fonètica (al·lòfon). Per veure el detall de la normalització *z-score* es pot observar l'esmentada normalització aplicada als subcostos en l'apartat 4.2.2.

La taula 4.1 mostra de manera detallada les estadístiques de primer ordre de cada paràmetre prosòdic estudiat. Si s'observa la taula resulta evident l'existència d'una lleugera descentralització de la mediana respecte la mitjana, cosa que permet constatar que les dades no segueixen una distribució normal.

A efectes d'analitzar la normalitat de les dades, a la taula 4.2 s'analitzen les estadístiques de segon ordre de cada paràmetre prosòdic. Observant les dades es pot confirmar l'asimetria de la distribució de les dades sobretot en el ritme i el *pitch*. Contraintuïtivament, es pot veure com la derivada del ritme presenta la millor simetria. Si s'analitza la forma de la distribució mitjançant el càlcul de la curtosi de les dades, es pot observar que el ritme i la derivada de F_0 presenten una concentració de les dades al voltant de la seva mitjana (presenten poca variació de dades). Per una banda, aquest resultat ens indica que el corpus té un ritme de producció de la parla constant i, per una altra, que no presenta variacions brusques en el contorn entonatiu. Si en comptes de disposar d'un corpus neutre s'analitzés un corpus expressiu amb emocions, aquests paràmetres adoptarien uns valors diferents en funció de l'emoció (generalment major variabilitat / menor concentració) (Iriondo, 2008) .

En canvi, un valor negatiu de curtosi en l'energia ens indica que les dades no es concentren al voltant de cap valor sinó que es reparteixen al llarg de tota la distribució. Per donar una explicació d'aquest valors, s'analitzen els histogrames de l'energia i la seva



(a) Histograma de la distribució de la energia a través de tot el corpus.

(b) Histograma de la distribució de la Δ d'energia a través de tot el corpus.

Figura 4.3: Histogrames de la prosòdia i la seva derivada per les energies en el corpus *url_fer_ct*.

derivada (figures 4.3(a) i 4.3(b)) on s'hi pot veure una concentració de les dades en dos regions diferents (distribució bimodal), degut a que els al·lòfons majoritàriament es poden dividir entre sonors i sords. Com a últim pas, s'analitza la normalitat de les distribucions dels paràmetres prosòdics dins del corpus mitjançant el test de Lilliefors (Lilliefors, 1967) (veure explicació detallada en l'apartat A.2 de l'annex A). Mitjançant el test de Lilliefors, s'obté un test d'hipòtesi que permet acceptar o rebutjar l'hipòtesi nul·la (H_0) que les dades s'adeqüen a una distribució normal. El valor llindar d'aquest estadístic per tal de decidir si la distribució s'adequa o no a la distribució normal depèn de dos factors: *i*) la mida de la mostra i *ii*) un nivell de significança α dins l'interval ([0.01-0.2] de menys a més significança). Segons s'indica (Lilliefors, 1967), per mostres $N > 30$ el valor llindar ve donat per l'expressió ($D_{max} = \frac{1.031}{\sqrt{N}}$).

Per tant, pel fet de disposar de 12,086 fonemes, el valor mínim d'acceptació per a la prosòdia és $D < 0.00937$ amb una $\alpha = 1\%$. En quant a la seva derivada, es disposa d'una mostra de 10565 valors, per tant el valor crític per a la derivada és $D < 0.01$. Si es miren els resultats de la taula 4.2 es pot observar que cap paràmetre prosòdic (ni tampoc les seves derivades) presenten una distribució normal, essent les derivades de l'energia i el ritme les que més s'hi acosten mentre que els valors de *pitch* són menys normals. En les figures C.1 i C.2 de l'annex C es poden veure amb detall els histogrames i *qqplots* (correlació de les dades respecte la normal) dels diferents paràmetres prosòdics del corpus.

4.2.2 Densitat dels subcostos i la seva normalització

Un cop descrita una primera anàlisi estadística de la prosòdia del corpus, s'analitza estadísticament la distribució dels diferents subcostos acústics que es consideraran en el càlcul de la funció de cost. Per a fer-ho d'una manera exhaustiva, s'analitzen per separat els subcostos d'unitat *target* i els subcostos de concatenació. A continuació s'explica l'anàlisi per a cada subcost.

Subcostos de *target*

Tal i com s'esmenta en l'apartat 2.3.3, els subcostos de *target* s'obtenen mitjançant el càlcul de la distància que hi ha entre els paràmetres acústics de les diferents unitats candidates, considerant la premissa que l'especificació generada pel mòdul de prosòdia es pot emular amb la prosòdia obtinguda de l'etiquetat de qualsevol unitat del corpus, servint així d'entrada per la funció de cost (Campbell i Black, 1997). En aquesta tesi doctoral s'assumeix que la prosòdia que guia la selecció de les unitats és ideal. És a dir, es considera com a prosòdia d'entrada els mateixos valors obtinguts de l'etiquetatge prosòdic del corpus (analitzats en l'apartat 4.2.1). Mitjançant aquesta assumpció es considera que independitzant la selecció d'unitats de la predicció prosòdica es poden dissociar els problemes per així fer-ne una optimització independent, sense que els errors d'un afectin l'altre (veure apartats 2.3.4 i 3.3.3).

Els subcostos de *target* amb que es treballa són: *i*) el *pitch* mig (en Hz), *ii*) l'energia (en RMS), i el *iii*) la durada de la unitat (en ms.). El *pitch* i l'energia presenten una continuïtat suprasegmental (a excepció dels sons sords) en la parla. En canvi, la durada, a diferència del *pitch*, l'energia o el ritme, és un paràmetre segmental específic per cada semial·lòfon (Navas *et al.*, 2002*b*). Per aquest motiu es separa el subcost de durada en dos subcostos: durada del semial·lòfon esquerre i durada del semial·lòfon dret. Els subcosts de *pitch* i energia es calculen de manera independent per cada semial·lòfon que conforma la unitat i s'amitjanen *posteriori* per obtenir el subcost d'unitat.

Llavors, s'obté el subcost d'unitat a través de la mitjana dels valors diferencials obtingut en els diferents semifonemes que conformen la unitat (en cas de difonemes 2 valors i en cas de trifonemes 3 valors). Els subcost de *pitch* només es té en compte pels semifonemes sonors ja que en els semifonemes sords no té sentit. La notació que es seguirà en aquesta tesi es PIT.T pel subcost de *pitch*, ENE.T pel d'energia, DURL.T pel de durada esquerra i DURR.T pel de durada dreta de la unitat, on T representa *target*.

Així doncs, les estadístiques referents als subcostos de *target* s'obtenen de manera in-

	Max	Min	Mitjana	Mediana	Desviació típica
PIT.T (Hz)	100.108	0	16.015	12.3329	13.5244
ENE.T (RMS)	0.2465	0	0.0311	0.0243	0.0268
DURL.T (ms.)	90.4582	0	10.6601	7.9515	10.1418
DURR.T (ms.)	121.379	0	11.4177	8.125	11.5923

Taula 4.3: Estadístiques de primer ordre dels subcostos de *target* del corpus *url_fer_ct*.

	Asimetria (*skewness*)	Curtosis	Test de Lilliefors
PIT.T	1.2604	1.5904	0.1078
ENE.T	1.4536	2.7238	0.1064
DURL.T	1.9367	5.6533	0.1500
DURR.T	2.1376	6.3631	0.1630

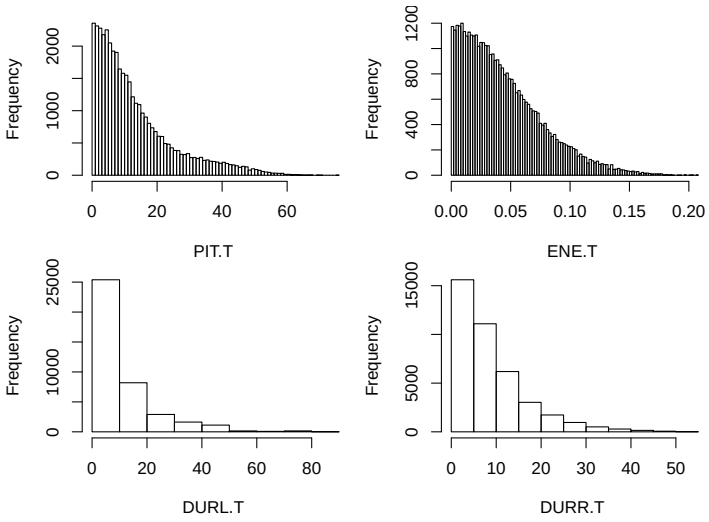
Taula 4.4: Estadístiques de segon ordre i proves de normalitat dels subcostos de *target* del corpus *url_fer_ct* (cap d'elles passa el test de Lilliefors).

dependent per a cada unitat: per a cada versió U_v d'una unitat U enregistrada en el corpus s'introdueix en la matriu d'anàlisi els subcostos derivats de canviar-la per la resta de versions U_{vv} (unitats que tenen la mateixa transcripció fonètica) mantenint l'especificació prosòdica original. D'aquesta manera s'obté una matriu $\binom{N}{2} \times M_T$ on N és el nombre de versions de la unitat i M_T el nombre de subcostos que s'analitzen ($M_T = 4$).

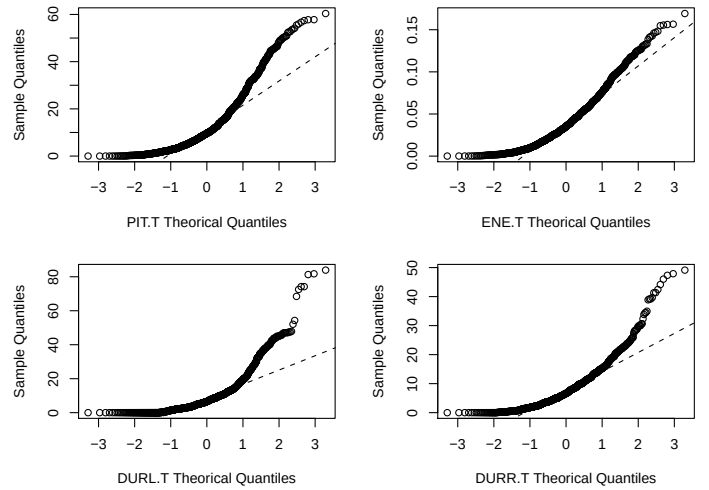
Les estadístiques de primer ordre obtingudes es mostren a la taula 4.3.

A tenor dels resultats de la taula, s'observa la necessitat de normalitzar els subcostos a un fons d'escala comú [0,1] per tal que puguin ser comparats entre si. Una altra vegada es pot intuir una asimetria en la distribució de dades pel fet de tenir valors diferents de mediana i mitjana alhora que aquests valors estan molt més propers al mínim del seu fons d'escala que del seu màxim. En quant a les desviacions típiques, aquestes adopten gairebé la mateixa magnitud que la mitjana, fet lògic considerant l'asimetria de les distribucions que indica que el valor que prenen la majoria de subcostos dins la funció de cost és en l'interval [0,μ] aproximadament.

Analitzant les estadístiques de segon ordre (taula 4.4), es confirma l'asimetria de les dades (*skewness* elevat) alhora que s'observa una alta concentració (curtosi) de dades en totes les distribucions. També es pot veure com els subcostos de durada són els que menys s'adeqüen a la distribució normal, ja que presenten els valors d'asimetria i curtosi més elevats. Si s'aplica el test de normalitat de Lilliefors (on el llindar per la mida de la mostra és $D < 0.0086$) es veu que cap subcost passa el test de normalitat alhora que es confir-



(a) Histogrames dels diferents subcostos de *target*.



(b) Comparació quartil-quartil dels subcostos de *target* respecte la distribució normal.

Figura 4.4: Histogrames i comparació quartil-quartil (*qqplot*) dels subcostos de *target* en el corpus *url_fer_ct*.

ma que els valors dels subcostos de durada són els que menys segueixen una distribució normal. A la figura 4.4(a) es poden veure els histogrames detallats per a cada subcost. Addicionalment es pot veure la comparació quartil-quartil a la figura 4.4(b).

Subcostos de concatenació

Anàlogament als subcostos de *target*, els subcostos de concatenació s'obtenen mitjançant el càlcul de la diferència de paràmetres acústics en el punt de concatenació (extrem) de la unitat (veure apartat 2.3.3).

Seguint de la prova de viabilitat (Alías, 2006), els subcostos de concatenació escollits són: *i*) la discontuinitat de *pitch* (Hz), *ii*) d'energia (RMS), i *iii*) de parametrització mel-cepstral (valor MFCC). El subcost MFCC es calcula com la distància dels 12 primers coeficients Mel-cepstrals en el punt de concatenació (exceptuant el C_0, que és l'energia) juntament amb les seves derivades (Campbell i Black, 1997). En les concatenacions sordes no es calcula el subcost de *pitch*, cosa que ja passava amb els subcostos de *target*. La notació seguida pels subcostos de concatenació és PIT.C per la discontinuïtat de *pitch*, ENE.C per la d'energia i MFC.C per la discontinuitat cepstral.

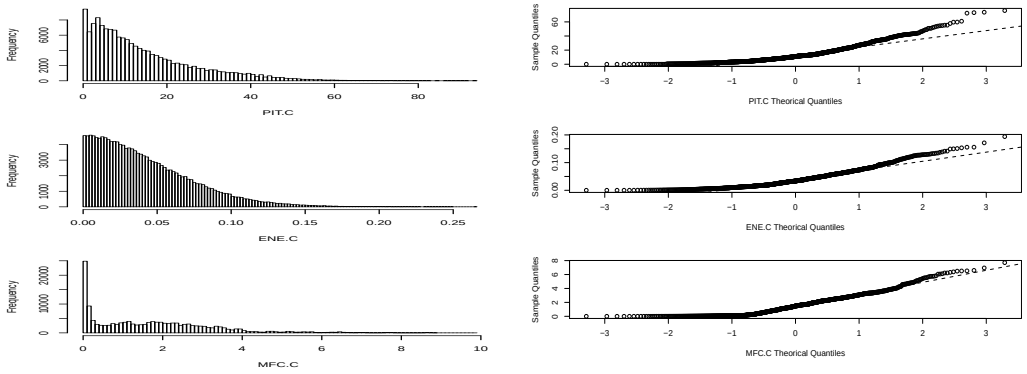
En el cas dels subcostos de concatenació, les estadístiques també s'obtenen de manera independent per a cada unitat (en el nostre cas la unitat anterior a la concatenació): per cadascuna de les variants U_i d'una unitat U del corpus es calculen els subcostos derivats de concatenar-la amb totes les possibles variants V_i on V representa totes les unitats que poden ser concatenades amb U (el semifonema dret d'U és el mateix que el semifonema esquerra d'V). Obtenint així una matriu $N' \times M_C$ on N' és el nombre de concatenacions possibles i M_C és el nombre de paràmetres analitzats ($M_C = 3$). Les estadístiques de primer ordre es mostren a la taula 4.5.

	Max	Min	Mitjana	Mediana	Desviació típica
PIT.C (Hz)	150.213	0	17.0496	13.5089	13.961
ENE.C (rms)	0.3772	0	0.0437	0.0321	0.0398
MFC.C (MFCC)	10.8608	0	1.7498	1.5803	1.5422

Taula 4.5: Estadístiques de primer ordre dels subcostos de concatenació del corpus *url_fer_ct*.

De nou, s'observa la necessitat de normalitzar els subcostos a un fons d'escala [0,1] per tal que els subcostos puguin ser comparats entre sí. Es torna a intuir una asimetria en

la distribució de dades al tenir valors diferents de mediana i mitjana. Tanmateix aquests valors són molt més pròxims al mínim del seu fons d'escala que del seu màxim. En quant a les desviacions típiques, aquestes adopten gairebé la mateixa magnitud que la mediana.



(a) Histogrames dels diferents subcostos de concatenació.

(b) Comparació quartil-quartil dels subcostos de concatenació respecte la distribució normal.

Figura 4.5: Histogrames i comparació quartil-quartil (*qqplot*) dels subcostos de concatenació en el corpus *url_fer_ct*.

	Asimetria (*skewness*)	Curtosis	Test de Lilliefors
PIT.C	1.1626	1.3526	0.1112
ENE.C	1.3972	2.1957	0.1135
MFC.C	0.8465	0.4499	0.1248

Taula 4.6: Estadístiques de segon ordre i proves de normalitat dels subcostos de concatenació del corpus *url_fer_ct* (cap d'elles passa el test de Lilliefors).

Analitzant les estadístiques de segon ordre (taula 4.6), s'observen dos grups clarament diferenciants en quant els valors d'asimetria i curtosi. Per una banda, s'observa que el *pitch* i l'energia presenten valors similars entre ells i que, en canvi, el subcost de MFC.C presenta una distribució més normal. Si s'aplica el test de Lilliefors s'observa que cap distribució passa el llindar (en aquest cas, $D < 0.0018$). Tot i així es pot observar com el subcost de MFC.C és el que s'acosta més a una distribució normal i el subcost d'ENE.C el que menys normalitat presenta. En quant a l'asimetria i la concentració de les dades s'observa que els subcostos de *pitch* i energia presenten semblança als valors dels mateixos paràmetres

fets servir pels subcostos de *target*. Altrament, es pot observar com el subcost MFC.C presenta uns valors d'asimetria i curtosi més baixos. A la figura 4.5(a) es poden veure els histogrames dels diferents subcostos de concatenació. Addicionalment a la figura 4.5(b) es pot veure la comparació quartil-quartil.

Normalització dels subcostos

L'anàlisi estadística dels subcostos evidencia la necessitat de normalitzar les dades a un fons d'escala comú dins l'interval [0,1] per a que puguin ser comparades entre elles (tal com s'havia explicat en l'apartat 2.3.5). D'altra banda al fer l'entrenament dels pesos mapant les distàncies cepstrals amb els subcostos mitjançant regressió lineal (veure apartat 2.4.2), els subcostos han d'acomplir certs requeriments tals com *i*) linealitat, *ii*) normalitat o *iii*) estabilitat de la variància (Chatterjee i Hadi, 2006), cosa que fa necessària la seva transformació.

A efectes de normalitzar el fons d'escala de les dades s'usen típicament dos mètodes: *i*) la normalització *z-score* (equació 4.1) i *ii*) la normalització *max-min* (equació 4.2) (Navas *et al.*, 2002*b*).

$$SC_i^{\text{Z-SCORED}} = \frac{SC_i - \overline{SC}}{\sigma_{SC}} \tag{4.1}$$

$$SC_i^{\text{MAX-MIN}} = \frac{SC_i - \min(SC)}{\max(SC) - \min(SC)} \tag{4.2}$$

$$\tag{4.3}$$

La normalització de dades, entesa en sentit estricte, no altera les estadístiques de segon ordre de la distribució (asimetria, curtosi i normalitat). Altrament, sí que modifica les estadístiques de primer ordre (fons d'escala, mediana, mitjana i desviació típica). El que marca la diferència entre les dues és com les modifiquen. Mentre que la normalització *z-score* força una distribució de mitjana 0 i desviació típica 1, no és capaç de fitar un fons d'escala determinat cosa que resulta molt contraproduent en el cas de la presència d'*outliers* en els extrems de la distribució. En canvi, si no es vol disposar de valors negatius (per definició els subcostos negatius no tenen sentit) la normalització *max-min* és capaç de fitar el fons d'escala en l'interval [0,1] sense fixar cap valor per la mitjana, la mediana o la desviació típica. Tot i així, pel fet de no canviar les estadístiques de segon ordre, la normalització *max-min* també és sensible als *outliers* com la normalització *z-score* (Jain *et al.*, 2005).

Per a poder obtenir optimitzacions mitjançant models lineals s'ha d'accentuar la normalitat de les distribucions canviant les estadístiques de segon ordre (Chatterjee i Hadi, 2006). Per tant, resulta necessari un procés de transformació de les distribucions (Tukey, 1957). Entre d'altres efectes, aquestes transformacions pretenen una millora de la simetria de la distribució, juntament amb la seva curtosi. Tal com s'ha explicat (apartat 4.2.1), el test de Lilliefors (Lilliefors, 1967) permet obtenir un índex (D) de normalitat d'una distribució de dades. Si aquest índex és inferior a un determinat llindar (calculat a partir de la mida de la distribució) llavors es pot assumir que la distribució de les dades segueix una distribució normal. En canvi, si aquest índex és superior al llindar es dedueix que les dades no segueixen una distribució normal. En el cas de comparar dues distribucions d'igual mida, aquest índex permet analitzar quina de les dues distribucions presenta més normalitat. Aquesta comparativa servirà per estudiar la normalitat de les diferents funcions d'aquest apartat. A l'apartat A.2 de l'annex A s'explica el principi de funcionament d'aquest test.

En la prova de viabilitat que representa el punt de partida d'aquesta tesi (Alías, 2006), s'aplica la transformació sigmoide (equació 4.4) emprada prèviament en la selecció d'unitats per (Febrer, 2001). A continuació s'analitzen detalladament els efectes d'aquesta transformació.

$$SC_i^{\text{SIGMOID}^2} = 1 - e^{-\left(\frac{SC_i}{\sigma_{SC}}\right)^2} \tag{4.4}$$

La transformació sigmoide es calcula per cada subcost segons la seva σ_{SC_u} especifica, evitant emprar una σ_{SC} global per a tot el corpus. L'esmentada transformació, canvia el fons d'escala de la distribució a l'interval [0,1], estabilitzant alhora la variància de la distribució de manera que els subcostos originals ubicats en l'interval [0,2σ] s'ubiquin en l'escala [0,0.9817], i els que son superiors a 2σ (inclosos en els *outliers*) estiguin en el tram (0.9817,1]. A la figura 4.6 es pot observar la funció de transformació sigmoide aplicada al subcost PIT.T. En l'esmentada figura també es pot veure com en més del 60% del fons d'escala el cost satura el seu valor a ≈ 1.

Per tal d'analitzar els efectes de la funció sigmoide sobre les dades originals, es tornen a obtenir les estadístiques de segon ordre un cop aplicada la transformació sobre tots els subcostos de *target* i concatenació (taula 4.7).

Analitzant els efectes d'aplicar la normalització sigmoide es poden observar impactes diferents en les estadístiques dels subcostos. A grans trets s'observa *i*) una estabilització de la desviació típica malgrat un augment de la seva magnitud ja que es passa d'una desviació típica de $\sigma_{\text{MAX}-\text{MIN}} = 0.1899 \pm 0.1396$ a $\sigma_{\text{SIGMOID}^2} = 0.3793 \pm 0.0154$ (si es descarta el subcost MFC.C per la seva diferent naturalesa, es passa de $\sigma_{\text{MAX}-\text{MIN}} = 0.1525 \pm 0.1082$

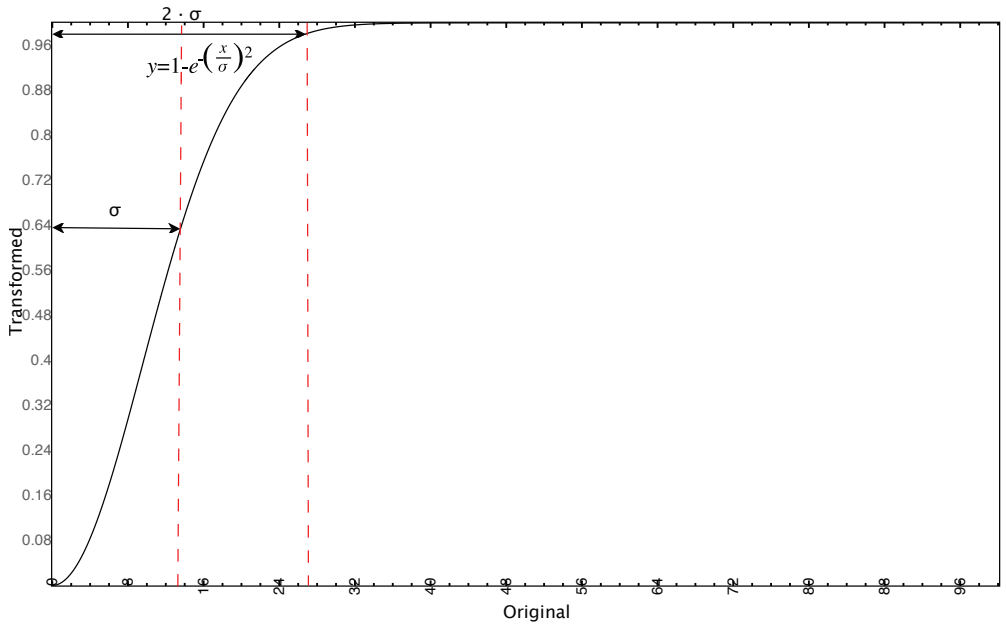


Figura 4.6: Funció de transformació sigmoide clàssica aplicada sobre el subcost PIT.T juntament amb la seva variant lineal.

	Desviació típica	Asimetria (*skewness*)	Curtosi	Test de Lilliefors
PIT.T	0.3743 *(0.1351)*	-0.1997 *(1.2604)*	-1.549 *(1.5904)*	0.1422 *(0.1078)*
ENE.T	0.3694 *(0.1087)*	-0.3343 *(1.4536)*	-1.463 *(2.7238)*	0.1450 *(0.1064)*
DURL.T	0.3755 *(0.1121)*	0.0014 *(1.9367)*	-1.5784 *(5.6533)*	0.1259 *(0.1500)*
DURR.T	0.3757 *(0.0955)*	0.0285 *(2.1376)*	-1.5761 *(6.3631)*	0.1162 *(0.1630)*
PIT.C	0.3749 *(0.0929)*	-0.2293 *(1.1626)*	-1.5422 *(1.3526)*	0.1376 *(0.1112)*
ENE.C	0.3713 *(0.1055)*	-0.2398 *(1.3972)*	-1.5202 *(2.1957)*	0.1470 *(0.1135)*
MFC.C	0.4137 *(0.1420)*	-0.2101 *(0.8465)*	-1.698 *(0.4499)*	0.1644 *(0.1248)*

Taula 4.7: Estadístiques de segon ordre obtingudes després d'aplicar la transformació sigmoide en els subcostos del corpus *url_fer_ct* (entre parèntesi i en cursiva es detalla el valor que s'obtindria amb una normalització *max-min*).

a $\sigma_{\text{SIGMOID}^2} = 0.3735 \pm 0.0026$), *ii*) una millora destacable de l'asimetria tot i que en molts casos comporta tenir una asimetria negativa i *iii*) una lleugera millora (millora en 5 dels 7 subcostos) de la baixa concentració de les dades tot i que també s'observa una major concentració de valors al voltant d'1 (queda reflectit amb una curtosi també negativa). En últim terme, malgrat la millora d'aquests valors, *iv*) si s'aplica el test de normalitat als subcostos es pot veure que només en 2 dels 7 subcostos (els referents a la durada) s'aconsegueix millorar l'estadístic D de normalitat dels subcostos i cap d'ells passa la prova.

És per aquest motiu que, en aquest treball de recerca es proposa, suavitzar la funció sigmoide clàssica amb una variant exponencial (a efectes de comparació l'anomenarem sigmoide lineal) que es detalla en l'equació 4.5:

$$SC_i^{\text{SIGMOID}} = 1 - e^{-\left(\frac{SC_i}{\sigma_{SC}}\right)} \tag{4.5}$$

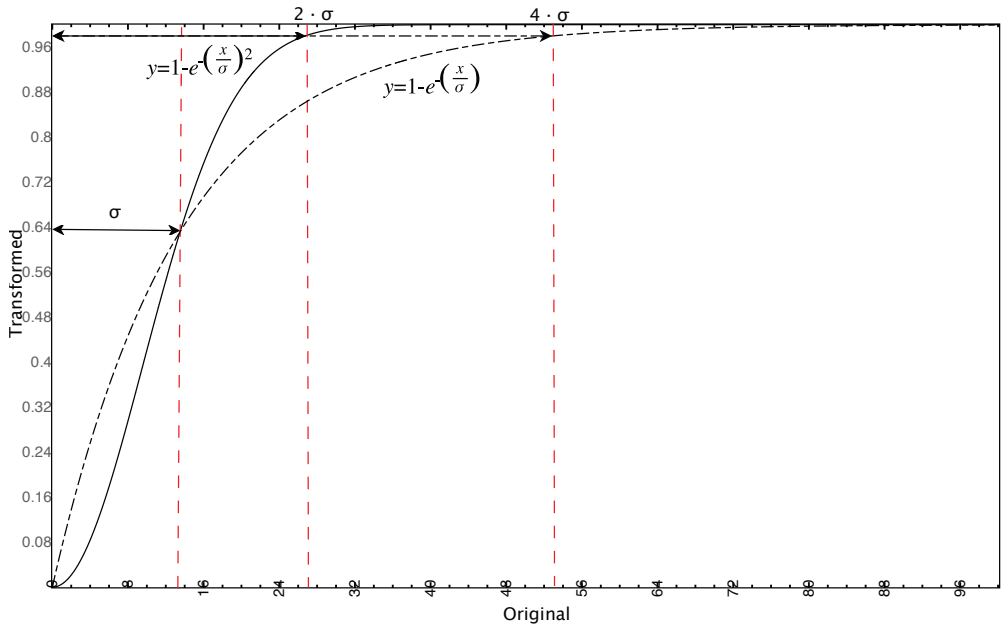


Figura 4.7: Funció de transformació sigmoide lineal (línia discontínua) aplicada sobre el subcost PIT.T comparada amb la funció sigmoide clàssica (línia sòlida).

En la figura 4.7 s'observa el comportament de la nova funció de normalització sigmoide lineal comparada amb la funció sigmoide clàssica. Amb detall, es pot veure que els efectes de transformació són molt semblants, però a diferència de la funció de sigmoide, la saturació dels subcostos ≈ 1 no comença a 2σ sinó que comença a 4σ, evitant així una saturació prematura del marge de valors. Llavors, es tornen a calcular les estadístiques

dels subcostos normalitzats mitjançant la taula 4.8.

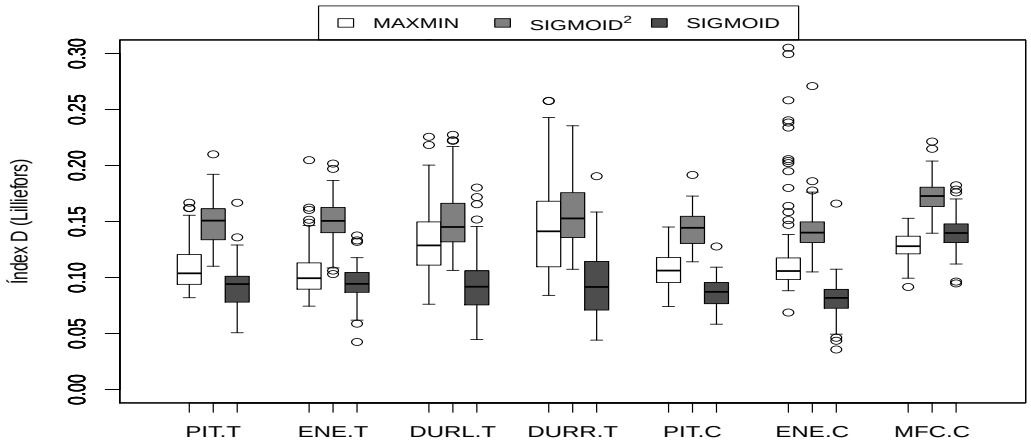
	Desviació típica	Asimetria (*skewness*)	Curtosi	Test de Lilliefors
PIT.T	0.2815	-0.3634	-1.0364	0.0828
ENE.T	0.2762	-0.4793	-0.9118	0.0886
DURL.T	0.2915	-0.2582	-1.0577	0.0551
DURR.T	0.292	-0.2267	-1.0789	0.0614
PIT.C	0.282	-0.394	-1.019	0.0808
ENE.C	0.278	-0.4016	-0.9892	0.0883
MFC.C	0.3392	-0.373	-1.4237	0.1317

Taula 4.8: Estadístiques de segon ordre obtingudes després d'aplicar transformació sigmoide lineal en els subcostos del corpus *url_fer_ct*.

Observant les estadístiques de nou, es pot veure que la desviació típica es desestabilitza lleugerament respecte la normalització clàssica passant a ser $\sigma_{\text{SIGMOID}} = 0.2915 \pm 0.0219$ ($\sigma_{\text{SIGMOID}} = 0.2836 \pm 0.0067$ sense MFC.C). Tot i això encara es més estable si es compara amb la normalització sense transformació (*max-min*). En quant a l'asimetria, també s'observa un lleuger empitjorament així com una lleugera millora en la curtosi (ja que no hi ha tants valors concentrats a ≈ 1). Però si s'analitza la normalitat de les dades, es pot veure que en tots els subcostos s'aconsegueix millorar el seu estadístic D assegurant que les dades, malgrat no ser normals, s'adeqüen millor a la distribució normal amb la transformació sigmoide lineal que amb la transformació sigmoide clàssica, doncs la transformació sigmoide clàssica malgrat millorar l'estabilitat de la desviació típica i els estadístics de segon ordre, empitjora la normalitat de les dades (veure taula 4.7).

Després d'estudiar la normalitat de la transformació dels subcostos considerant les distribucions que adopten aquests en les diferents unitats, s'estudia el comportament de la transformació per a cada unitat específica, considerant less 100 unitats més poblades del corpus. Llavors, per cadascuna d'aquestes unitats es mira el seu índex de normalitat D de Lilliefors segons *i*) la normalització sense transformació (*max-min*), *ii*) la transformació sigmoide clàssica i *iii*) la transformació sigmoide lineal. Els estadístics D obtinguts es mostren a la figura 4.8(a) i el valor mig obtingut (mitjana del valor D per cada subcost a través de les 100 unitats més poblades del corpus) es mostra en la taula de la figura 4.8.

En aquesta última anàlisi, es confirmen, a nivell d'unitat, els resultats obtinguts a nivell global: l'index de normalitat D empitjora si s'aplica la transformació sigmoide clàssica



(a) *Boxplots*

mitjana	MAX-MIN	$SIGMOID^2$	$SIGMOID$
PIT.T	0.109	0.1479	0.0913
ENE.T	0.1042	0.1512	0.0941
DURL.T	0.1343	0.1509	0.0937
DURR.T	0.1434	0.1552	0.0939
PIT.C	0.107	0.1437	0.0869
ENE.C	0.1211	0.1405	0.0797
MFC.C	0.1282	0.1724	0.1397

(b) Taula

Figura 4.8: Resultat del test de Lilliefors aplicat als subcostos de les 100 unitat segons les normalitzacions max-min , sigmoide clàssica ($SIGMOID^2$) i sigmoide lineal ($SIGMOID$) detallats per (a) *boxplot* i (b) taula. La taula recull la mitjana de l'índex D a través de les 100 unitats més poblades del corpus *url_fer_ct*.

respecte les dades originals, en canvi la transformació sigmoide lineal millora l'estadístic de normalitat. No obstant això, en cap cas es passa el valor llindar i per tant les dades no segueixen una distribució normal.

Per tant, la funció de transformació sigmoide lineal és la que s'utilitzarà com a funció de normalització en el càlcul dels subcostos del corpus *url_fer_ct*, que es descriuen a continuació.

A l'annex C es detallen les gràfiques de les diferents transformacions aplicades en aquest apartat.

No obstant l'estudi realitzat hi ha un parell d'aspectes que cal observar de cara a les normalitzacions. El primer aspecte és que s'ha realitzat un estudi de transformacions clàssiques, homogènies per tot el domini de les dades originals deixant de banda altres transformacions més complexes que assoleixen plena normalitat aplicant transformacions heterogènies (varien la funció de transformació en funció del valor de les dades d'entrada – Saon *et al.* (2004)). El segon aspecte a destacar és que l'estudi de normalitat dels subcostos es necessari per l'ajust de pesos mitjançant mètodes de regressió lineal (MLR/NNLS) ja que no és imprescindible per la cerca de pesos mitjançant els mètodes evolutius.

4.3 Anàlisi de la fiabilitat dels mètodes d'ajust automàtic

Un cop estudiada la naturalesa del subcostos en el corpus de veu, es pot entrar a revisar amb detall la ponderació dels subcostos dins la funció de cost. Abans d'entrar d'entrar a l'anàlisi de l'ajust de pesos perceptiu mitjançant aiGA, es realitza una revisió del comportament de les diferents tècniques d'ajust automàtic així juntament amb l'anàlisi de la fiabilitat dels pesos obtinguts. Dotar d'un índex de fiabilitat l'ajust de pesos automàtic mitjançant distàncies cepstrals permet avaluar la metodologia independentment de la comparativa amb els mètodes d'ajust perceptius i a la vegada entendre millor les dificultats del mapatge d'aquest tipus de dades.

En aquest estudi de fiabilitat, s'estudien els ajustos de pesos proposats per (Alías, 2006): la regressió lineal (MLR - apartat 2.4.2) i l'optimització evolutiva (GA - apartat 3.2.4) juntament amb el seu grau de fiabilitat.

4.3.1 Fiabilitat dels patrons de pesos obtinguts amb MLR/NNLS

En l'apartat 2.4.2 s'ha explicat amb detall el funcionament de la metodologia d'ajust de pesos MLR mitjançant l'algorisme NNLS i les distàncies cepstrals. Addicionalment, en l'apartat 3.2.4 d'ha analitzat la naturalesa d'aquests resultats en termes de linealitat i *fitness* global. Tanmateix, no s'ha analitzat la fiabilitat d'aquests resultats mitjançant el càlcul de l'estadístic R^2 (ja explicat en l'apartat 2.4.2) o l'error quadràtic mig.

En aquest apartat, s'expandeix l'estudi realitzat a l'apartat 3.2.4 per veure quina fiabilitat tenen aquests pesos. A continuació es repassen de manera resumida els dos estadístics més típics per analitzar la fiabilitat d'una regressió lineal, ja explicats a l'apartat 2.4.2:

Error quadràtic mitjà RMSE:

L'error quadràtic mitjà (*Root Mean Squared Error* - RMSE) (Netter *et al.*, 1990) és la mitjana dels quadrats de l'error comès pel model de regressió lineal respecte a la distribució de dades original. Formalment es defineix segons:

$$\text{RMSE} = \frac{\sum_i^N e_i^2}{N} = \frac{\sum_i^N (y_i - \hat{y}_i)^2}{N} \tag{4.6}$$

on y_i és la distància cepstral real respecte la unitat d'origen i $\hat{y}_i$ és la predita a través de la ponderació particular dels subcostos segons uns pesos determinats. En general, interessa que aquest error tingui el valor més proper a zero possible, el que indicaria un ajust lineal perfecte a la mostra original.

Coeficient de determinació R^2:

Tal com s'ha dit a l'apartat 2.4.2, el coeficient de determinació indica la consistència de l'ajust mesurant la qualitat del model de regressió en funció de la desviació (percentatge) de dades que aquest pot explicar (Netter *et al.*, 1990). En altres paraules el coeficient de determinació mesura quina proporció de la variabilitat que presenten les dades és explicada pel model lineal. El model de determinació es formula de la manera següent:

$$R^2 = 1 - \frac{SS_{Error}}{SS_{Total}} \tag{4.7}$$

on SS_{Total} és la variància de les dades d'entrenament (distàncies cepstrals) i SS_{Error} és la variància dels errors o variància residual que coincideix amb:

$$\text{RMSE} = SS_{Error} \tag{4.8}$$

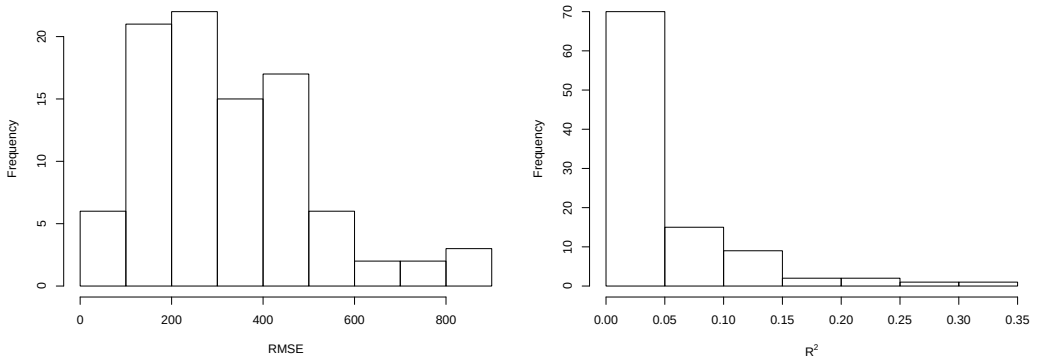
Com més proper d'1 estigui el coeficient de determinació, millor serà l'aproximació lineal. El coeficient de determinació coincideix amb el quadrat del coeficient de correlació:

$$R^2 = r_{xy}^2 \tag{4.9}$$

Si la correlació és negativa, $r_{xy} << 0$ la relació és inversa. En canvi, la correlació es positiva $r_{xy} >> 0$, la relació és directa.

Resultats

Si s'analitzen els resultats de R^2 i l'RMSE obtinguts en l'ajust dels pesos mitjançant MLR en les 100 unitats més poblades del corpus s'obtenen els resultats de la taula 4.9, mostrats com histogrames en la figura 4.9.



(a) RMSE (b) R^2

Figura 4.9: Histogrames dels estadístics RMSE i R^2 obtinguts en ajustar les 100 unitats més poblades del corpus *url_fer_ct* mitjançant MLR.

Estadística	Min	1Q	Mediana	Mitjana	3Q	Max
RMSE	70.5	198.3	308.2	383.4	467.1	2234
R^2	$4.4 \cdot 10^{-6}$%	0.8%	2.6%	4.6%	6.1%	30.8%

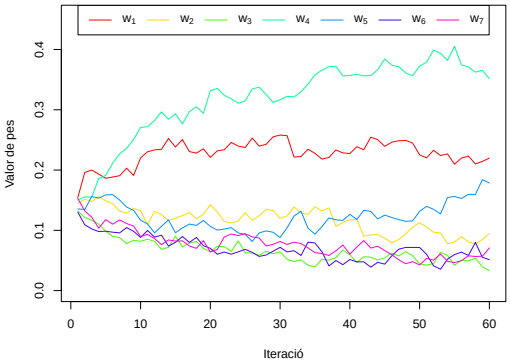
Taula 4.9: Estadístiques obtingudes pel MLR.

Analitzant els resultats $RMSE$, es pot observar que hi ha un valor mig de $RMSE$ molt alt ($\overline{RMSE} = 383.4$) quan la mitjana de les distàncies cepstrals es troba en 87.43 ± 28.3517. Altrament, es pot observar que si s'analitza el coeficient de determinació R^2 per als pesos obtinguts, la regressió només és capaç de modelar de mitjana el 4.6% dels subcostos respecte les distàncies cepstrals. Els índex R^2 i RMSE permeten concloure que els pesos obtinguts mitjançant regressió lineal no resulten fiables segons el propi model. Aquest fet indica un comportament no lineal dels subcostos escollits respecte les distàncies cepstrals que evidencien la necessitat d'optimitzar els pesos de manera no lineal.

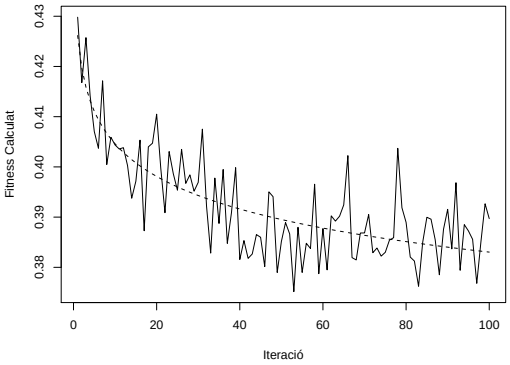
4.3.2 Fiabilitat dels patrons de pesos obtinguts amb GA

El algorismes genètics amb selecció per torneig presenten robustesa i consistència en entorns sorollosos tal com s'ha esmentat en l'apartat 3.2.4. En els GA clàssics es mesura la convergència de l'algorisme a través d'un indicador que mesura l'evolució del *fitness* mig a través de les generacions (Goldberg, 2002). Observant l'evolució d'aquest *fitness* en les 100 primeres generacions (figura 4.10(b)) es pot confirmar l'alta presència de soroll en l'ajust de pesos. El soroll, en un context evolutiu, es defineix com a canvis en la cerca produïts per l'atzar degut a una pobre representació genètica (Goldberg, 1989) . Tanmateix, es pot veure una corba de tendència de tipus exponencial que indica un lleugera convergència.

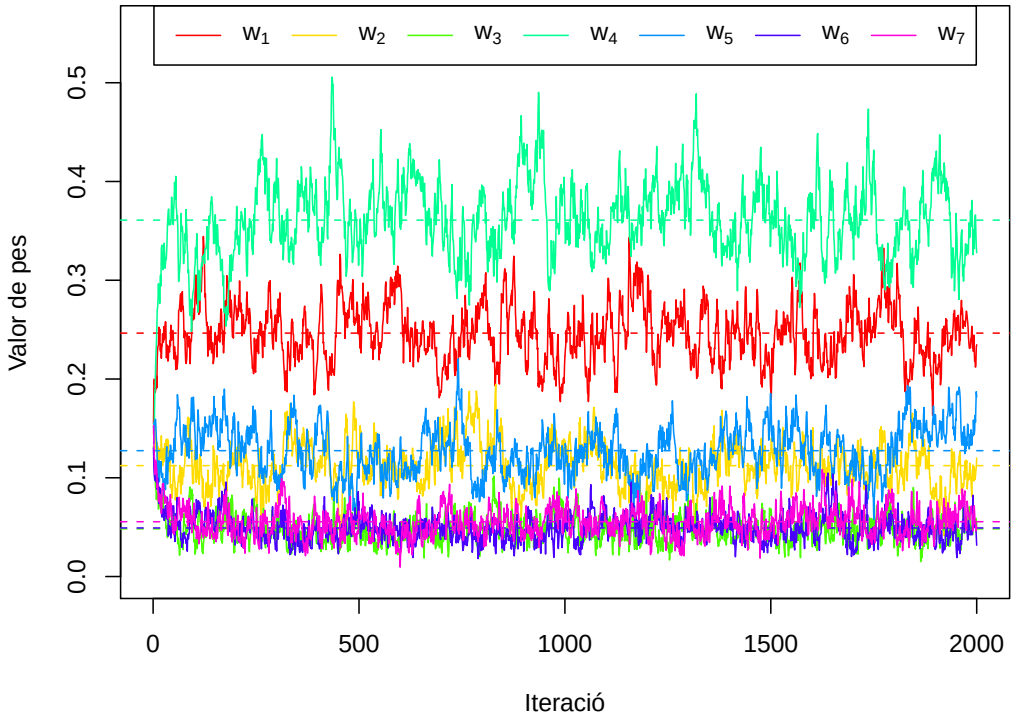
Tot i la convergència sorollosa del *fitness*, aquesta no garanteix del tot la consistència o robustesa de les dades. En aquest sentit, es proposa analitzar la variació del valor dels diferents valors de pesos al llarg del procés evolutiu. En la figura 4.10(a) s'observa l'evolució dels valors de pesos en les 100 primeres iteracions del procés mentre que en la figura 4.10(c) es pot veure l'evolució completa per una execució de 2000 iteracions. Malgrat que la figura 4.10(a) pot aparentar un comportament no sorollós en l'evolució, quan s'analitza la figura 4.10(c) es pot veure com els pesos presenten un comportament sorollós al voltant d'un valor mig. Una evolució consistent ve marcada per la progressió del *fitness* no sorollosa. Si la progressió del *fitness* no oscil·la degut al soroll, el % de variació del pes sobre la mitjana hauria de ser proper a 0, un cop s'hagués convergit. En canvi, si malgrat estabilitzar-se al voltant d'un valor mitjà el valor del pes segueix variant en un percentatge elevat, significa que la direcció de la cerca genètica es guia, en diferent mesura segons l'atzar. Seguint aquest raonament i tenint en compte l'índex de soroll que presenten els pesos s'avalua com a mesura de consistència el % de variació (desviació típica) de valor del pes respecte el seu valor mig. Els valors obtinguts es detallen en la taula i gràfica de la figura 4.11.



(a) Primeres 100 iteracions.

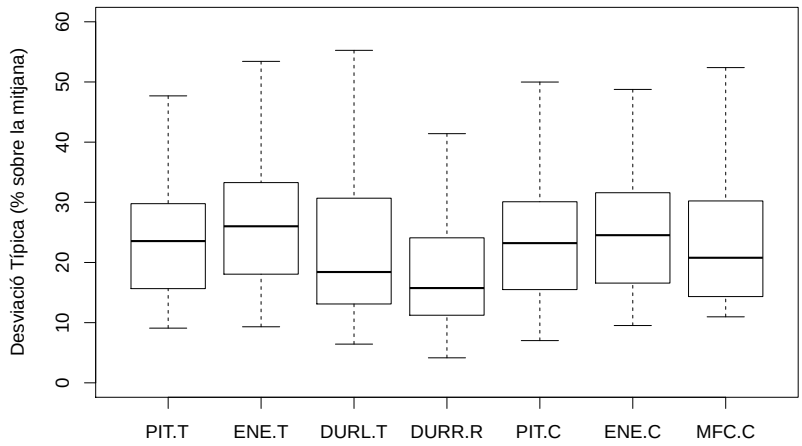


(b) Fitness 100 primeres iteracions.



(c) Execució completa (2000 iteracions).

Figura 4.10: Detall de l'evolució dels pesos per la unitat /@l/ que és la que té més representació en el corpus *url_fer_ct* . En les gràfiques w_1 =PIT.T, w_2 =ENE.T, w_3 =DURL.T, w_4 =DURR.T, w_5 =PIT.C, w_6 =ENE.C, w_7 =MFC.C.



Subcost	Min	1Q	Mediana	Mitjana	3Q	Max
PIT.T	9.09%	15.70%	23.56%	26.56%	29.74%	138.84%
ENE.T	9.315%	18.068%	26.017%	28.818%	33.201%	137.341%
DURL.T	6.434%	13.135%	18.416%	25.186%	30.476%	197.834%
DURR.T	4.157%	11.244%	15.751%	19.835%	24.076%	71.933%
PIT.C	7.02%	15.56%	23.22%	27.26	29.80%	200.68%
ENE.C	9.522%	16.635%	24.538%	27.895%	31.555%	136.595%
MFC.C	10.97%	14.38%	20.78%	24.59%	30.23%	70.21%

Figura 4.11: Desviacions típiques (en % sobre el valor de la mitjana) dels valors dels pesos en el transcurs de 2000 generacions en un GA.

Analitzant les dades obtingudes, es pot veure que en mitjana els pesos varien un 25% el seu valor en el transcurs del procés evolutiu. En l'apartat 4.3.1 s'ha destacat la poca o nul·la fiabilitat dels pesos obtinguts a través de MLR.

4.4 Precisió en el nivell d'ajust dels pesos

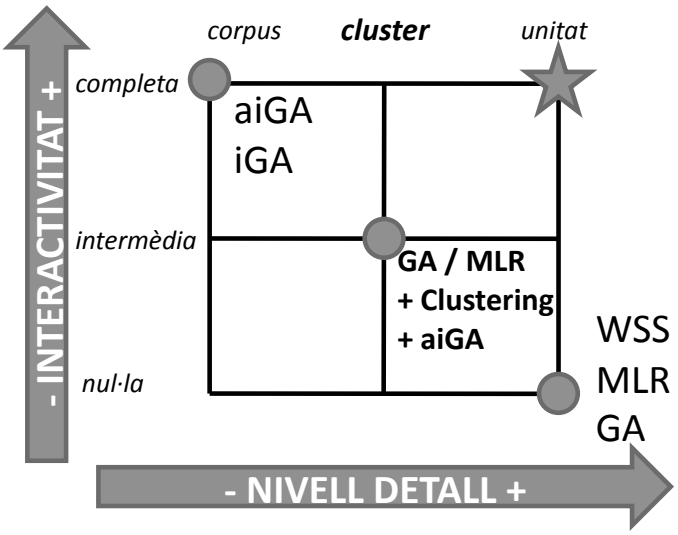


Figura 4.12: Nivells d'ajust de pesos possibles en funció de la interactivitat i precisió que ofereix el mètode d'ajust (l'estrella mostra l'ajust desitjat).

En la revisió de l'estat de la qüestió (apartat 2.4.2) s'ha explicat que el procés d'entrenament dels pesos es pot realitzar a dos nivells diferents: *i*) a nivell d'unitat (per a cada fonema del corpus de veu) o a nivell global (és a dir, per a les unitats de tot el corpus en el seu conjunt) (Hunt i Black, 1996; Black i Taylor, 1997*a*). Amb anterioritat a aquesta tesi, inclòs en el seu punt de partida (Alías, 2006), el nivell d'ajust ha vingut determinat per la dificultat tècnica del procés: els pesos s'ajustaven a nivell d'unitat quan l'ajust era objectiu i automàtic (MLR - apartat 2.4.2, GA - apartat 3.2.4), i a nivell global quan l'ajust es realitzava subjectivament mitjançant un conjunt molt reduït d'expressions a través de iGAs o aiGAs (apartats 3.3 i 3.4).

No obstant això, l'estat de l'art també constata la importància relativa dels subcostos en funció de la unitat (o tipus d'unitat) que es vol recuperar mitjançant la funció de cost (Campillo *et al.*, 2005). Tanmateix, realitzar l'ajust de pesos a nivell d'unitat de manera perceptiva resulta pràcticament impossible degut a que tant *i*) l'ajust d'una sola unitat

resulta gairebé imperceptible a l'usuari, com *ii*) el número de proves elevat que s'hauria de realitzar provoquen que l'esmentada aproximació no resulti factible.

En aquesta tesi, partint del treball de Colotte i Beaufort (2005) i Campillo *et al.* (2005), es proposa organitzar l'espai de cerca en grups (o clústers) de pesos en funció del seu comportament en les metodologies d'ajust automàtic: si l'especificitat acústica de diferents unitats implica un canvi en la importància dels subcostos per determinar les versions cepstralment més properes (Campillo *et al.*, 2005), aquest coneixement queda reflectit en els pesos obtinguts a l'hora de modelar aquesta importància. Per tant, es poden agrupar les unitats en funció del comportament (patró) dels pesos obtinguts en l'ajust automàtic i refinar el valor dels mateixos mitjançant l'ajust perceptiu (figura 4.13). És important remarcar que els clústers només determinen quins pesos s'han d'emprar (Colotte i Beaufort, 2005) i, per tant, no restringeixen la selecció d'unitats a unes unitats (versions) concretes en en front a unes altres, a diferència de Black i Taylor (1997*a*). En altres paraules la metodologia de detectar patrons de pesos permet evitar la generalització dels pesos provocat per ajustar unitats de diferent naturalesa (Campillo *et al.*, 2005) de manera conjunta (p.ex, vocal / consonant). El fet de realitzar l'ajust per a totes les unitats alhora implica no saber la contribució real de cada subcost a la qualitat sintètica final per cada unitat. Conseqüentment, l'aproximació mitjançant clústers permet obtenir diferents patrons de pesos per cada grup d'unitats que guarden similitud entre elles en temes d'importància de subcost. Llavors, s'assoleix un nivell de precisió intermedi (figura 4.12) entre l'ajust global (totes les unitats juntes) i l'ajust a nivell d'unitat (un pes per cada unitat) (Meron i Hirose, 1999) però que alhora respecta la ponderació dels subcostos en funció de l'especificitat de la unitat (Campillo *et al.*, 2005)

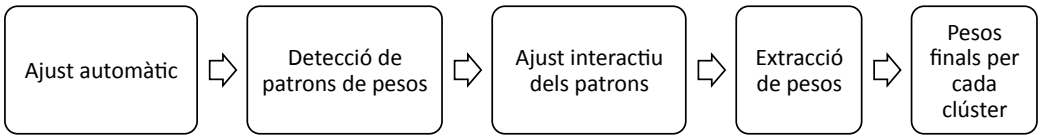


Figura 4.13: Diagrama de les etapes de la metodologia proposada per ajustar pesos a nivell de clúster.

Addicionalment, com a fet destacable, l'ajust de pesos mitjançant clústers, resulta assolible de posar-lo en pràctica. Superant així els inconvenients de fatiga i la complexitat de disseny que comportaria l'ajust per cada unitat de manera perceptiva.

A continuació, es detalla la complexitat que presenta realitzar l'ajust a nivell de clúster (p.ex., avaluar de manera perceptiva només unitats concretes d'una frase, impedint que la resta d'unitats influeixin en la decisió).

4.4.1 Ajust perceptiu dels patrons obtinguts

Un cop obtinguts diferents patrons de pesos (un per cada clúster), els clústers s'optimitzen perceptivament mitjançant aiGA (veure apartat 3.4). A tal efecte, en primer lloc, es seleccionen frases fonèticament balancejades del corpus que continguin majoritàriament unitats del clúster que s'està ajustant. Aquestes frases s'obtenen mitjançant un algorisme de cerca que obeeix a dos criteris: *i*) maximitzar el nombre d'unitats consecutives que pertanyen al clúster i *ii*) minimitzar l'entropia de clústers diferents presents en la frase. Després, es construeix un corpus ad hoc per a cada frase segons les següents premisses: *i*) les unitats no pertanyents al clúster optimitzat (en aquest treball denominades com unitats portadores) es fixen deixant només la versió corresponent a la frase del corpus que s'està optimitzant, i *ii*) les unitats que pertanyen al clúster que s'està optimitzant (unitats variables) aporten totes les seves versions per a ser seleccionades amb excepció de les versions pertanyents a la frase que s'està sintetitzant per a fer l'ajust (Black i Tokuda, 2005).

4.4.2 Prosòdia emprada en l'ajust perceptiu

En el procés perceptiu d'ajust de pesos es pren com a referència la prosòdia original extreta de les unitats del corpus de veu, procés conegut com a *copy-prosody* (Montero *et al.*, 1999), desacoblant així els errors de la predicció de prosòdia (veure apartat 4.2.2). No obstant això, la prosòdia original no està exempta d'errors degut a que els mètodes de segmentació i etiquetat no són perfectes. En aquest cas, però, la segmentació i etiquetat del corpus ha estat revisada manualment, permetent així que la prosòdia obtinguda no contingui errors. El fet de tenir una prosòdia ideal (obtinguda de la veu natural) permet disposar també d'un senyal de veu real que es pot mostrar als usuaris avaluadors com una referència prosòdica a seguir pel selector d'unitats. En canvi, si es treballés amb una prosòdia sintètica no es podria disposar de cap veu natural de referència en el procés d'ajust subjectiu.

4.4.3 Generació de forma d'ona en l'ajust perceptiu: WAV vs. TD-PSOLA

La qualitat de la selecció d'unitats és difícil de mesurar de manera aïllada ja que la síntesi depèn de tots els mòduls que composen un CTP-SU, tal com s'ha vist en la justificació del *copy-prosody*. De tota manera, el fet d'incloure el mòdul de composició de forma d'ona és una qüestió no resolta. Per exemple, alguns treballs advoquen per independitzar l'optimització dels diferents mòduls del CTP-SU (Hunt i Black, 1996) mentre que d'altres proposen la inclusió d'alguna mena de processament del senyal en el cicle d'ajust (Meron i Hirose,

1999) per ajustar el sistema en global, degut a la dependència dels diferents mòduls.

Tal com s'ha dit, s'ha optat per aïllar el problema de la selecció d'unitats de la predicció de prosòdia del sistema per tal de facilitar l'optimització del procés. Tal com s'ha explicat en l'apartat 4.2.2, pel fet d'emprar una prosòdia real extreta del corpus s'aconsegueix no propagar els problemes d'una predicció prosòdica, malgrat que en altres corpus pot tenir errors d'etiquetat. Seguint la mateixa línia, en aquest treball es proposa estudiar l'impacte de minimitzar el processament del senyal necessari per la composició de la forma d'ona. A tal efecte, un cop seleccionades les unitats, el senyal de les mateixes es recupera íntegrament del corpus sense realitzar-hi cap modificació prosòdica. Simplement, per concatenar dues unitats, es realitza un mínim ajustament de la fase per evitar discontinuitats de fase que puguin comportar problemes de concatenació (artefactes) en la síntesi final (veure apartat 2.2.4). No obstant, per a poder analitzar l'emmascarament que provoca el mòdul de composició de forma d'ona respecte el mòdul de selecció d'unitats, es realitzen els ajustos de pesos amb ambdues tècniques (concatenació directa (WAV) i concatenació TD-PSOLA amb modificacions prosòdiques).

4.4.4 Agrupació de pesos mitjançant arbres de classificació i regressió

En aquest apartat i com a aproximació inicial al problema, es parteix dels pesos obtinguts automàticament per obtenir els clústers en funció de l'especificitat fonètica. Entre les dues metodologies d'ajust de pesos automàtiques, es seleccionen els pesos obtinguts mitjançant GA (apartat 3.2.4). El motiu d'elecció d'aquests pesos és que els pesos obtinguts mitjançant MLR presenten un coeficient de determinació (R^2) i error quadràtic mig (RMSE) molt baix. De totes maneres, malgrat que els GA presentin certa estabilitat en la cerca, aquesta no és ideal (té un component sorollós). Com a factor addicional per aquesta elecció, cal sumar-hi el fet que en el treball previ realitzat per Alías i Llorà (2003) (apartat 3.2.4) els pesos obtinguts mitjançant GA obtenen un millor *fitness* objectiu que els pesos obtinguts mitjançant MLR.

Els pesos, s'agrupen segons els trets fonètics de les seves unitats emprant un arbre de classificació i regressió (CART) (s'usa l'eina *wagon* de la plataforma de Festival (Black i Taylor, 1997*b*)). Agrupant els pesos en funció de la tipificació fonètica de l'unitat s'asseguren clústers coherents sense limitar el nombre de versions de la unitat a seleccionar. En canvi, si es realitzés l'agrupació de les unitats (Black i Taylor, 1997*a*) sense tenir en compte els pesos de la mateixa (p.ex. distància cepstral acústica) no hi hauria cap garantia d'assolir coherència respecte el comportament dels pesos, a més esdevindria un cercle viciós

si aquesta distància objectiva es volgués ponderar tal i com es fa a Black i Taylor (1997*a*). D'altra banda, la metodologia CART tracta de manera implícita la dispersió d'unitats amb poques realitzacions (Black i Taylor, 1997*a*), obtenint la divisió fonètica que millor minimitza l'entropia de cada clúster i així, evitant l'agrupament de pesos d'unitats fonèticament diferents. El joc de preguntes del CART inclou la informació següent per a cada semifonema que conforma la unitat: *tipus* (vocal, consonant, semivocal o silenci), *sonoritat* (sonor o sord), el *mode d'articulació* (oclusiu, fricatiu, etc.) i el *lloc d'articulació* (bilabial, dental, etc). A efectes de calcular l'entropia, s'ha de proporcionar una matriu de distàncies a l'arbre de regressió. Per calcular la matriu de distàncies entre els diferents vectors de pesos es poden emprar diferents distàncies vectorials (cosinus, euclídea...) (Qian *et al.*, 2004). En el cas dels pesos, degut a que la funció de cost es normalitza segons la suma dels pesos (p.ex. $\sum w_i = 1$, veure l'equació (2.1)), es tria la distància del cosinus per a calcular la similitud vectorial entre dos vectors de pesos. Aquesta decisió es pren ja que la diferència entre dos vectors de pesos és l'angle geomètric entre ells. En canvi, la diferència de mòduls no resulta significativa a efectes d'agrupament (Qian *et al.*, 2004). Finalment, un cop s'ha obtingut l'arbre de decisió amb el màxim de clústers possible, es determina el número òptim de clústers analitzant l'impuresa dels clústers obtinguts, tal i com es descriu a continuació.

Nombre de clústers

L'objectiu d'aquest apartat és definir el procés emprat per a determinar el número de clústers òptims analitzant la impuresa de les dades dels clústers obtinguts. Com a primer pas, es construeix un arbre CART emprant com a dades els pesos de les 100 unitats més poblades del corpus, obtinguts mitjançant GA. Així, s'obté un conjunt de preguntes fonètiques en forma d'arbre que permet subdividir el corpus sencer. Després, es determina el nombre òptim de clústers. A tal efecte, es consideren indicadors de puresa dels clústers (Günter i Bunke, 2003) que analitzen aspectes com la compactació i separació dels clústers, la similitud mitja entre clústers propers entre sí, l'aïllament entre clústers i la coherència intra-clúster o la maximització de distàncies entre clústers minimitzant la distància intra-clúster. Concretament les mesures emprades són (Wang *et al.*, 2009): Silhouette, Davies-Bouldin, Calinski- Harabasz, Dunn, Hubert-Levin (C-*index*), Krzanowski-Lai i Hartigan (Dudoit i Fridlyand, 2002; Bolshakova i Azuaje, 2006), l'índex ponderat intra-inter clúster (*weighted inter-intra index*) (Strehl, 2002) i homogeneïtat (Sharan *et al.*, 2003).

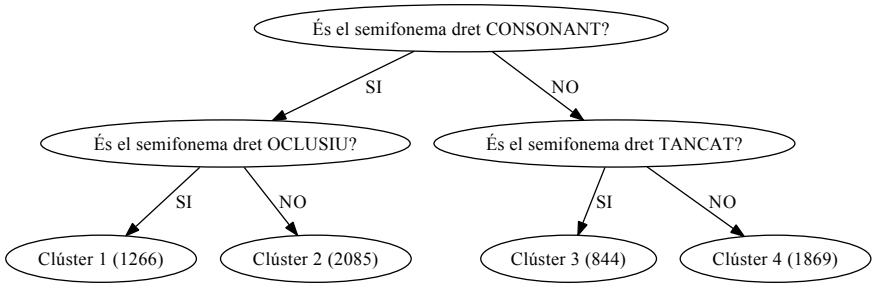
Indicador/#clusters	2	3	4	5
Silhouette	**0.1072**	0.0753	-0.0447	-0.0748
Davies-Bouldin	**1.896**	3.4578	7.5295	7.5091
Calinski-Harabasz	13.114	**24.458**	17.361	14.584
Dunn	**0.8272**	0.2835	0.1072	0.113
C-index	0.4201	**0.349**	0.3545	0.3771
Krzanowski-Lai	0.4547	**3.0914**	2.9821	0.3346
Hartigan	13.114	31.695	**2.4406**	4.4043
weighted inter/intra	0.2281	**0.5348**	0.4734	0.4702
Homogeneity	0.628	**0.6692**	0.6649	0.6474

Taula 4.10: Resultat de les mesures d'impuresa de *clustering* emprades per determinar el nombre òptim de clústers. En negreta el valor òptim per a cada mesura.

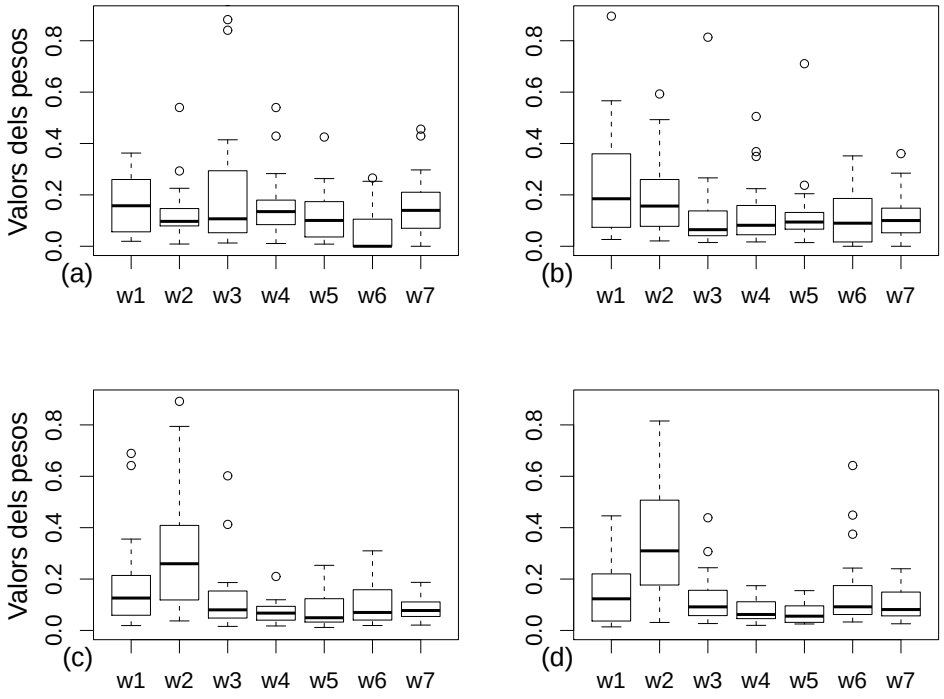
La taula 4.10 mostra que els diferents indicadors determinen el nombre de clústers òptim entre 2 i 4 clústers. Després d'analitzar la distribució de les unitats (mirar el nombre d'unitats a cada clúster de la figura 4.14(a)), es selecciona 4 com el número de clústers final que a ajustar perceptualment. Malgrat que el clúster 3 i el clúster 4 representen patrons de pesos objectius bastant similars (*boxplots* de la figura 4.14(b)), el número significativament gran de realitzacions d'unitats contingudes en cas de fusionar els clústers, ocasionaria una partició de dades massa desequilibrada (1266, 2085, 2713 vs. 1266,2085,844,1869), fet que implicaria una manca de precisió a l'hora de respectar les especificitats característiques de les diferents unitats. Cal afegir, tanmateix, que 4 clústers encara resulta un nombre factible considerant el nombre de proves subjectives a realitzar.

4.5 Definició de nous indicadors per l'aiGA

Anàlogament al que s'ha explicat a l'apartat 4.3 un dels elements clau en l'extracció de resultats del procés d'ajust és la robustesa dels pesos obtinguts. A efectes d'avaluar el model perceptiu obtingut, resulta necessari disposar d'indicadors que mesurin la bondat del procés evolutiu. En l'apartat 3.4.7, s'ha descrit un primer indicador κ per mesurar la consistència dels usuaris (veure l'equació 3.5). La mesura de consistència (κ) considera cicles ($A > B$, $B > C$ i $C > A$) dins el graf com contradiccions, provocant una disminució de la consistència del model. Aquesta mesura es computa com la proporció entre els vertexs en cicles i tots els vèrtexs dins del graf (veure equació 3.5). Així, es poden descartar aque-



(a) CART obtingut limitant a 4 el número de clústers. El nombre d'elements per cada clúster s'indica entre parèntesi.



(b) *Boxplots* dels valors dels pesos obtinguts després d'agrupar els patrons de pesos basats en GA en (a) clúster 1, (b) clúster 2, (c) clúster 3 i (d) clúster 4. Cal dir que w_1 s'empra per a indicar el pes de DURL.T, w_2 per a DURR.T, w_3 per a ENE.C, w_4 per a ENE.T, w_5 per a MFC.C, w_6 per a PIT.C i w_7 per a PIT.T.

Figura 4.14: Arbre de decisió amb el conjunt de preguntes emprat per clusteritzar els pesos obtinguts usant GA. Les gràfiques de la subfigura (b) mostren els patrons de pesos obtinguts per cada clúster.

lles evolucions provinents de les respostes contradictòries dels usuaris abans de procedir a l'extracció de resultats.

Seguint aquest enfocament, en aquest capítol s'expandeixen la definició d'indicadors del procés evolutiu específics per l'aiGA amb els següents objectius: *i*) poder descartar execucions que no hagin convergit o bé difereixin dels altres usuaris avaluadors a nivell de la solució obtinguda, i *ii*) poder determinar el nombre d'iteracions vàlides a causa d'un comportament sorollós per part de l'usuari, és a dir, el procés de construcció del graf es veu obligat a aturar-se prematurament bé perque l'usuari ja ha arribat a una solució òptima o bé perquè s'ha fatigat. A continuació, es presenten els tres nous indicadors definits en el marc d'aquest treball d'investigació (Formiga *et al.*, 2010):

4.5.1 Índex de certesa λ

Aquesta mesura dóna informació sobre el grau de confusió dins del graf. Anàlogament a l'índex de consistència, aquesta mesura considera les comparacions avaluades com empats ($A = B$) com un augment de l'ambigüitat en la cerca evolutiva. Aquesta ambigüitat pot ser deguda a diferents factors, tals com la convergència prematura de la població o la fatiga de l'avaluador. Així, la mesura es defineix com l'invers (1 menys) del quocient entre el nombre de vèrtexs que relacionen empats i el nombre total de vèrtexs dins del graf. Aquesta mesura permet obtenir informació sobre la durada eficaç de la prova perceptiva (fins on aporta informació seguir construint el graf). Una cop la prova evolutiva convergeix, les síntesis candidates presentades en l'avaluació poden resultar indistingibles a l'usuari i per tant les etiqueta amb empats. Llavors convé interrompre el graf prematurament. L'indicador es defineix en l'equació 4.10.

$$\lambda\left(\mathcal{G}'^{t}, \omega_\lambda\right) = 1 - \left(\frac{1}{\left|\mathcal{V}'^{t}\right|} \cdot \sum_{v \in \psi(\mathcal{G}'^{t})} \omega_v^\lambda\right)^{\alpha_\lambda} \tag{4.10}$$

on $\left|\mathcal{V}'^{t}\right|$ es el nombre de vèrtex de $\mathcal{G}'$ en l'instant t, ω_v^λ és el pes del vèrtex v en la mesura, $\psi\left(\mathcal{G}'^{t}\right)$ representa els vèrtex en empats detectats en $\mathcal{G}'^{t}$, i α_λ un factor d'escalat global superior o igual a 1. Si no es diu el contrari, la mesura es calcula considerant $\omega_v^\lambda = 1, \forall v \in \mathcal{V}'^{t}$ i $\alpha_i = 1$ (Formiga *et al.*, 2010).

4.5.2 Índex de convergència intra-usuari ρ

Aquesta mesura dóna informació sobre la convergència de la prova perceptiva cap a una o múltiples solucions per a un mateix usuari. Donat que l'aiGA no inclou cap esquema de substitució en el procés evolutiu, la solució final no s'obté a partir de la població sencera sinó que s'extreu d'aquells individus més ben classificats (de manera heurística es consideren el 10% d'individus millors en la classificació $\hat{r}(v)$). Tanmateix, no hi ha cap garantia que els individus més ben classificats hagin convergit a un únic valor (configuració de pesos per a un mateix usuari), cosa que implica considerar un indicador que permeti mesurar aquest efecte. La convergència intra-usuari es calcula com la mitjana de la matriu de correlacions (correlació de cosinus) de les solucions millor classificades (veure l'equació 4.11).

$$\rho\left(\mathcal{G}'^{t}\right) = \left(\frac{1}{|\mathcal{B}'^{t}|} \cdot \sum_{\substack{\forall v \in \mathcal{B}'^{t} \\ \forall v' \in \left\{\mathcal{B}'^{t} | v' \neq v\right\}}} corr(v, v') \right)^{\alpha_\rho} \tag{4.11}$$

on $\mathcal{B}'^{t}$ són els millors vèrtexs a $\mathcal{G}'$ en l'instant t per un sol usuari, $\left|\mathcal{B}'^{t}\right|$ és la mida de $\mathcal{B}'^{t}$ (10% millors), α_ρ un factor d'escalat global superior o igual a 1 i $corr(v, v')$ denota la correlació lineal mitjançant distància del cosinus entre les configuracions de pesos v i v' que pertanyen a $\mathcal{B}'^{t}$. Si no es diu el contrari, $\alpha_\rho = 1$ (Formiga *et al.*, 2010).

4.5.3 Índex de correlació inter-usuari τ

Aquesta mesura dóna informació sobre la similitud de proves perceptives realitzades per diferents usuaris. Per a cada prova perceptiva, es considera els individus més ben classificats (10%) anàlogament al que s'ha considerat per l'índex de convergència intra-usuari. Llavors, aquestes solucions es comparen entre elles a través d'una nova matriu de correlacions inter-usuari. De nou, l'indicador final s'obté mitjançant la mitjana de la matriu de correlacions. La mesura es detalla en l'equació 4.12.

$$\tau\left(\mathcal{G}'^{t}\right) = \left(\frac{1}{|\mathcal{U}'^{t}|} \cdot \sum_{\substack{\forall v \in \mathcal{U}'^{t} \\ \forall vv \in \{\mathcal{U}'^{t} | vv \neq v\}}} corr(v, vv) \right)^{\alpha_\tau} \tag{4.12}$$

on $\mathcal{U}'^{t}$ són els millors vèrtexs en $\mathcal{G}'$ en l'instant t per tots els usuaris, $\left|\mathcal{U}'^{t}\right|$ és la mida de $\mathcal{U}'^{t}$ (millors pesos de tots els usuaris), α_τ un factor d'escalat global superior o igual a 1 i $corr(v, vv)$ denota la correlació lineal mitjançant distància del cosinus entre les configuracions de pesos v i vv que pertanyen a $\mathcal{U}'^{t}$. Si no es diu el contrari, $\alpha_\tau = 1$ (Formiga *et al.*, 2010).

4.6 Ajust perceptiu dels pesos mitjançant aiGA

4.6.1 Disseny de l'ajust a nivell de clúster

En els experiments que es presenten a continuació, l'ajust perceptiu es presenta als usuaris mitjançant la plataforma WTISS (Formiga, 2003). Perseguint l'objectiu d'evitar avaluacions excessivament ambigües (apartat 3.4.2) s'adapta la interfície per a facilitar l'avaluació de l'usuari. Donat que les frases sintetitzades contenen unitats fixes i unitats variables (apartat 4.4.1) , es proporciona el text de la frase sintetitzada subratllant les unitats variables, cosa que permet que l'usuari pugui centrar la seva atenció en les diferències de les unitats variables de les diferents síntesis presentades.

A l'hora d'incorporar els clústers de l'apartat 4.4.4 al procés, s'elegeixen 16 frases representatives (4 per clúster) del corpus mitjançant l'algorisme d'entropia esmentat en l'apartat 4.4.1 (taula 4.11). A nivell estadístic les frases contenen 30.2 ± 12.6 unitats, de les quals 11.6 ± 5.23 són variables i seleccionables mitjançant la funció de cost de selecció. A cada usuari se li demana que ajusti perceptivament 4 frases d'almenys dos clústers diferents. En total 21 usuaris van participar en les proves garantint que cada frase fos ajustada per 4.8 usuaris de mitjana. Cada execució de l'aiGA amb l'usuari dura de 13 ± 3 minuts (considerant que es podien escoltar els fitxers tants cops com fos necessari).

Clúster	Frase	Unitats Variables	Unitats Totals
1	Enll***oc de cada d***ia.	4	13
1	Té e***n c****ompte* que e***l ca***talà en general.	9	25
1	I s***ob***retot *d*els me***mb***re***s d******el p******art it d***emòc***rat***a.	12	28
1	E***n q***uina lle***ng***ua ha***n p***arl***at t***o***ts p******leg******at***s.	8	25
2	Seg***ons les*** dad***es*** de l'aj***unt******ament***.	11	22
2	***I s***obretot d***els mem***br***es*** d***el*** partit d***em***òcrata.	10	28
2	***Alguns fenòmens***.	6	11
2	***Aix******òés el*** que ***n***o agrada ***als més ra******cialmen***t ***espanyols***.	16	35
3	***Bil******ingüi***sme selec***tiu i a******i***llaments.	6	26
3	Se***ria ins******tructiu i po***tser e***dif******i***cant.	9	27
3	***Cons******titucio*** del parlament per ele***gi***r el ***mini***stre ***princi***pal del ***go***vern exe***cutiu i*** el s***eu*** ad***ju***nt.	21	59
3	***T******i***ngui com ***a o***bjec***tiu u***n ***vint-i-ci***nc de pel·***lícu***les d'è***xit d******o***blades al català.	13	50
4	***Que ha estat a***l ***Barça en aquestes darreres setmane***s.	16	30
4	***No falta la raó*** al con***seller Hernà***n***dez***.	12	26
4	El ***nombre de*** ciclis***te***s ur***ba***ns ***també ha a***ug***menta***t ***molt e***ls ***ca***p ***de setmana***	23	47
4	Els botxins ***d'Espanya, e******limina***ts ***de***l mundi***al***	10	32

Taula 4.11: Frases escollides per realitzar les proves interactives usant aiGA dissenyades pel corpus *url_fer_ct*. Amb negreta es destaquen les unitats variables (que admeten selecció d'unitats) respecte les unitats portadores (fixes en tot l'ajust).

Seguint la mateixa posada a punt que en l'apartat 3.4.3 les generacions evolutives es composen de 16 combinacions de pesos comparades segons un torneig binari (15 comparacions entre 2 mostres de veu sintètiques a cada iteració). El procés evolutiu continua durant 4 generacions. En tot moment la consistència, ambigüitat, convergència i correlació dels usuaris es controla per mitjà dels indicadors κ, λ, ρ i τ (apartats 3.4.7 i 4.5).

Per a poder comparar els efectes de l'ajust mitjançant síntesi WAV i síntesi TD-PSOLA (apartat 4.4.3), cada usuari ajusta 3 proves mitjançant síntesi WAV i 1 prova mitjançant TD-PSOLA.

4.6.2 Extracció de resultats

Un cop finalitzat el procés d'ajust, les avaluacions poc robustes de cada prova (amb consistència $\kappa < 1$, segons l'equació 3.5) es descarten per a les anàlisis posteriors. Un cop considerats només els resultats robustos, es determina si l'usuari ha proporcionat informació significativa fins a l'última iteració o no. Tal com s'ha esmentat, això pot ocórrer bé perquè s'ha trobat una solució òptima abans de finalitzar els tornejos, o bé perquè l'usuari s'ha fatigat, sent incapaç d'adonar-se de les variacions presents en l'última iteració del procés evolutiu donada la seva subtilesa. El criteri d'aturada del procés iteratiu es determina, doncs, considerant l'indicador λ explicat en l'apartat 4.5. Es considera l'última iteració on l'usuari augmenta l'índex de certesa com el punt on l'usuari ha convergit (es considera que més enllà d'aquest punt hi ha una tendència a la disminució de la certesa fins al final del procés evolutiu – veure figura 4.15). El procés es detalla formalment a continuació:

$$\begin{aligned} \Lambda &= \left[\lambda'_1, \ldots, \lambda'_t, \ldots, \lambda'_N\right], \ on \\ \lambda'_t &= \frac{\lambda(\mathcal{G}'^{t-1}) + \lambda(\mathcal{G}'^{t}) + \lambda(\mathcal{G}'^{t+1})}{3} \end{aligned}$$

$$\delta\lambda'_t = \lambda'_{t+1} - \lambda'_t, \forall t \in [1, \ldots, N-1] \tag{4.13}$$

$$it_{FINAL} = \text{últim}(\delta\lambda'_t \geq 0) \tag{4.14}$$

on λ'_t és el valor mig de $\{\lambda(\mathcal{G}'^{t-1}), \lambda(\mathcal{G}'^{t}), \lambda(\mathcal{G}'^{t+1})\}$ de la mètrica λ del graf $\mathcal{G}'$ a l'iteració t. N és el nombre total d'avaluacions realitzades per l'usuari. Λ és el conjunt de λs normalitzades essent $\delta\lambda'_t$ la seva derivada (pendent).

La idea per obtenir la iteració (it_{FINAL}) de cada usuari és la següent: es calcula l'índex $\lambda(\mathcal{G}'^t)$ per cada etapa (instant t) de construcció (avaluació de l'usuari) del graf $\mathcal{G}'^t$ normalitzat (sense empats). Llavors, per reduir les oscil·lacions locals, els valors obtinguts d'aquest indicador es suavitzen amitjanant-los amb el seu valor anterior $\lambda(\mathcal{G}'^{t-1})$ i posterior $\lambda(\mathcal{G}'^{t+1})$ obtenint així l'index λ'_t suavitzat. Llavors, s'obté la derivada $\delta\lambda'_t$ de la mètrica λ'_t través de la diferència del seu valor λ'_t amb el seu valor immediatament posterior en el temps λ'_{t+1}. Al final, és considera la ultima iteració bona (it_{FINAL}) com aquella més tardana on la derivada és positiva (augmenta la certesa). A la figura 4.15 s'hi pot veure un cas pràctic.

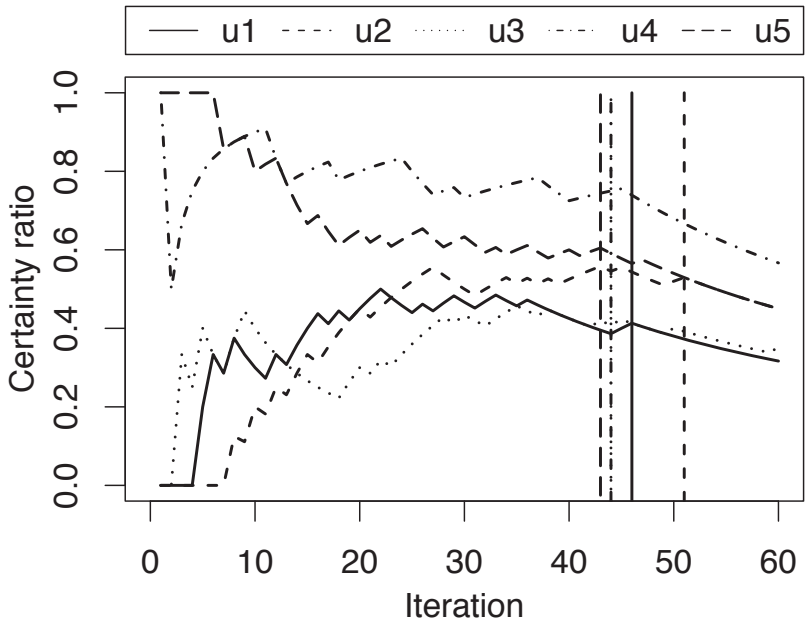


Figura 4.15: Evolució de la certesa (λ) de cinc usuaris ($u_1, \ldots, u_5$) a través del procés evolutiu per una frase particular del clúster 1. Les línies verticals de la gràfica indiquen l'última iteració significativa en termes de certesa per cadascun dels 5 usuaris.

Un cop s'han construït els models basats en el graf per a tot el clúster considerant les duples <usuari,frase> s'obté un patró de pesos concret per cada clúster que integri les solucions on han arribat els diferents usuaris (patró de consens).

L'aproximació més senzilla a aquest problema és trobar el vector de pesos real més proper al centroide del clúster que conformen els millors pesos de les duples. La idea de treballar amb un vector de pesos real i no un amitjanat ve donada pel fet que la configuració de pesos ha d'haver estat sintetitzada i avaluada. Un vector de pesos amitjanat que no s'hagi avaluat perceptivament no garanteix realitzar una bona síntesi. El procés per

obtenir el vector de consens és el següent.

Per cada dupla es consideren els seus millors vèrtexs com a solucions òptimes. En aquest cas són els vèrtexs ubicats en el 10% de les millors posicions del *ranking* $\hat{r}(v)$. Llavors, aquests vèrtexs es seleccionen com les solucions on ha convergit l'usuari per la frase en qüestió. El genotip d'aquests vèrtexs s'afegeix a una matriu de solucions multiusuari W^c. Aquesta matriu representa els millors pesos de tots els usuaris. La matriu W^c és de dimensions $N \times M$ on N és el nombre de vèrtexs i M els diferents pesos (M=7 en el nostre cas). Llavors es calcula la mediana per cada columna de la matriu per tal d'obtenir el vector de pesos final w^c, que representa el patró al que s'ha convergit. S'empra la mediana en comptes de fer la mitjana ja que es treballa amb valors discretitzats amb dos decimals i així es pot obtenir un valor que existeix realment. Per exemple, si els valors de pesos obtinguts fossin $\{0.02, 0.10, 0.10, 0.11\}$ s'observaria que el valor 0.02 és un *outlier* en la distribució, per tant valor mig quedaria millor representat per la mediana 0.10 en comptes de la mitjana 0.08, que no indica cap valor real en concret.

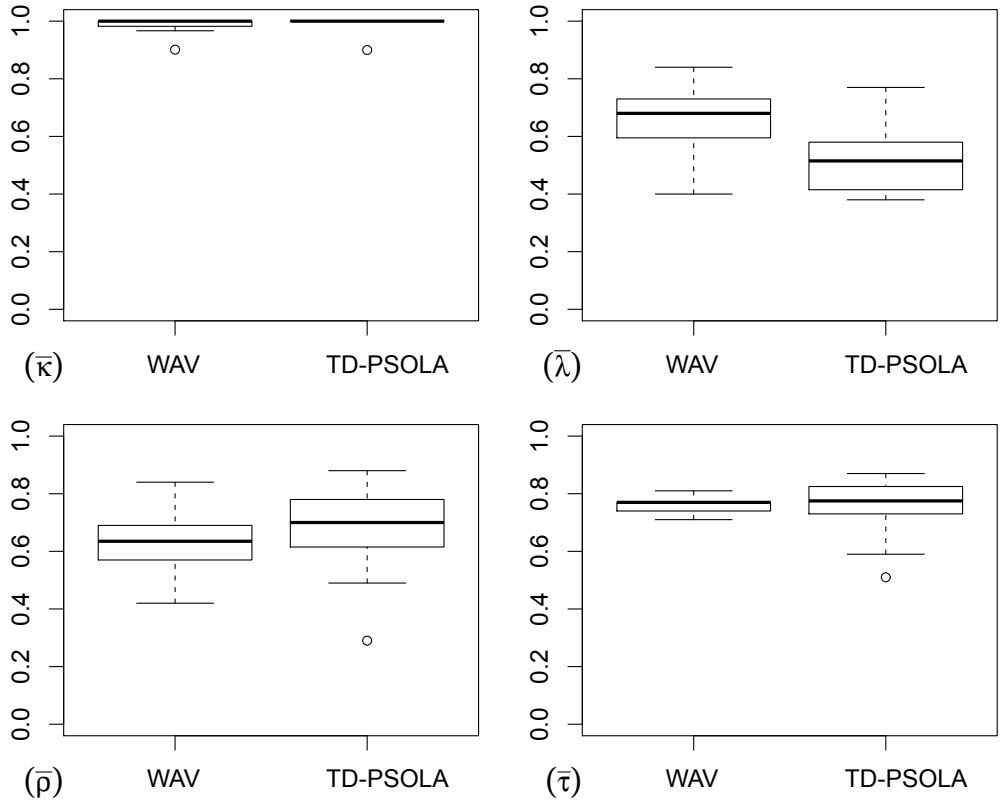
Com a últim pas, es normalitza el patró de pesos ($\sum w_i^c = 1 : \forall w_i^c \in w^c$) i es busca la configuració de pesos W_i ja existent dins W^c que minimitza la distància del cosinus amb w^c. El vector de pesos seleccionat W_i és el que es considera com a millor combinació de pesos per un clúster determinat. Una altra manera d'entendre aquesta tècnica és realitzar una cerca al veí més proper (1-*Nearest Neighbour* o 1-NN) al centroide dels millors pesos, calculant aquest centroide a través de la mediana i no la mitjana.

4.6.3 Resultats

A l'hora d'analitzar les dades obtingudes, primer s'estudien els resultats dels indicadors obtinguts segons els dos mètodes de síntesi emprats (WAV / TD-PSOLA) per determinar la fiabilitat dels pesos i posteriorment, un cop triada una configuració vàlida, s'analitzen els pesos obtinguts comparant els seus pesos a nivell de clúster amb els pesos obtinguts segons MLR i GA (a nivell d'unitat).

Comparativa WAV / TD-PSOLA

En la figura 4.16 es poden observar els resultats de l'indicador de consistència κ juntament amb els nous indicadors λ, ρ i τ proposats. Els resultats es detallen per a cada tècnica de síntesi (WAV / TD-PSOLA) emprada, i per les 16 frases (4 per clúster) que han participat en l'ajust perceptiu realitzat per tots els usuaris que han participat en l'ajust per cada frase.



Frase	$\bar{\kappa}$		$\bar{\lambda}$		$\bar{\rho}$		$\bar{\tau}$	
	WAV	TD-PSOLA	WAV	TD-PSOLA	WAV	TD-PSOLA	WAV	TD-PSOLA
1	0.9	**1**	**0.54**	0.52	0.63	**0.75**	**0.78**	0.73
2	**1**	**1**	**0.6**	0.38	0.42	**0.63**	0.71	**0.78**
3	**1**	**1**	**0.73**	0.56	0.64	**0.81**	0.8	**0.82**
4	0.97	**1**	0.67	**0.71**	0.56	**0.67**	0.74	**0.84**
5	0.97	**1**	**0.84**	0.46	**0.7**	0.67	**0.77**	**0.77**
6	0.98	**1**	**0.61**	0.51	**0.84**	0.74	**0.81**	0.51
7	**1**	0.9	**0.81**	0.6	0.67	**0.84**	0.77	**0.87**
8	0.98	**1**	0.68	**0.69**	0.66	**0.73**	0.74	**0.86**
9	**1**	**1**	0.59	**0.77**	**0.68**	0.49	**0.77**	0.65
10	**1**	**1**	**0.68**	0.44	0.71	**0.86**	0.77	**0.83**
11	**1**	**1**	0.5	**0.56**	0.46	**0.6**	**0.73**	**0.73**
12	**1**	**1**	**0.78**	0.56	0.58	**0.73**	0.75	**0.8**
13	**1**	**1**	**0.7**	0.46	0.6	**0.64**	0.77	**0.78**
14	0.99	**1**	**0.73**	0.38	**0.74**	0.29	**0.77**	0.59
15	**1**	**1**	**0.4**	0.39	0.48	**0.88**	**0.77**	**0.77**
16	**1**	**1**	**0.73**	0.39	**0.62**	0.56	0.74	**0.75**

Figura 4.16: Resultats dels diferents indicadors segons el mètode de síntesi (en negreta la millor tècnica en cada cas).

Per a corroborar les diferències que puguin haver-hi entre les dues tècniques de síntesi considerades en els valors dels indicadors, es realitzen proves de significança per cada parella d'indicadors. Per a cada parella es realitza una prova t de Student (Leon-Garcia, 1994). A partir dels resultats d'aquesta anàlisi de significància estadística s'obtenen les conclusions següents: *i*) $\kappa_{WAV} \approx \kappa_{TD-PSOLA}$ (p-value=0.4589), *ii*) $\lambda_{WAV} > \lambda_{TD-PSOLA}$ (p-value=0.0028), *iii*) $\rho_{WAV} \approx \rho_{TD-PSOLA}$ (p-value=0.2353), *iv*) $\tau_{WAV} \approx \tau_{TD-PSOLA}$ (p-value=0.7882).

L'indicador de certesa demostra que introduir processament del senyal en el mòdul de composició de forma d'ona emmascara l'ajust de pesos interactiu. En augmentar la sensació d'igualtat en les diferents variants de síntesi presentades a l'usuari, aquest introdueix ambigüitat en la cerca provocant un índex de certesa que és significativament inferior ($\lambda_{TD-PSOLA} = 0.52 \pm 0.12$) en comparació a l'ajust de la selecció d'unitats amb un mínim processament del senyal $\lambda_{WAV} = 0.66 \pm 0.11$. No obstant això, convé destacar que l'ajust mitjançant TD-PSOLA no ha provocat un augment en les contradiccions del usuaris ($\kappa_{TD-PSOLA} = 0.99 \pm 0.02$ i $\kappa_{WAV} = 0.99 \pm 0.02$) ja que aquests han adoptat una actitud conservadora en front a la igualtat. En aquest sentit és confirma la hipòtesi que l'índex de consistència κ resultava insuficient per avaluar la qualitat de les solucions obtingudes.

A nivell de genotip, analitzant els valors dels pesos obtinguts, s'observa que l'excés d'ambigüitat no afecta a la convergència (ρ) o a les correlacions inter-usuari (τ), demostrant que l'aiGA és robust cercar combinacions de pesos en entorns ambigus o sorollosos.

No obstant això, i vist l'excés d'ambigüitat que el processament del senyal introdueix en el procés interactiu, es considera treballar amb els pesos obtinguts amb la composició d'ona WAV seguint l'objectiu de desacoblar el problema de la selecció d'unitats de la resta de mòduls del TTS. En aquesta decisió s'assumeix que es realitza una simplificació del problema aïllant els pesos i la funció de cost de la tècnica de síntesi emprada en la composició de la forma d'ona (com també ho eren dels errors en la predicció prosòdica).

Consistència, convergència i correlació dels pesos

La figura 4.17 mostra un exemple de la mesura de consistència de 5 usuaris per a una frase particular al llarg dels tornejos. En la figura es pot veure com hi ha tres usuaris que són sempre coherents, un d'ells (*u1*) pot recuperar la consistència gràcies a la modelització dels aiGA basada en grafs, mentre que l'*u4* és incapaç de retornar a una consistència $\kappa = 1$ a causa de les diferents caigudes de consistència significatives que presenta. Com a resultat, es descarten 6 de les 77 (7.8%) duples de les parelles avaluador/frase ja que el procés

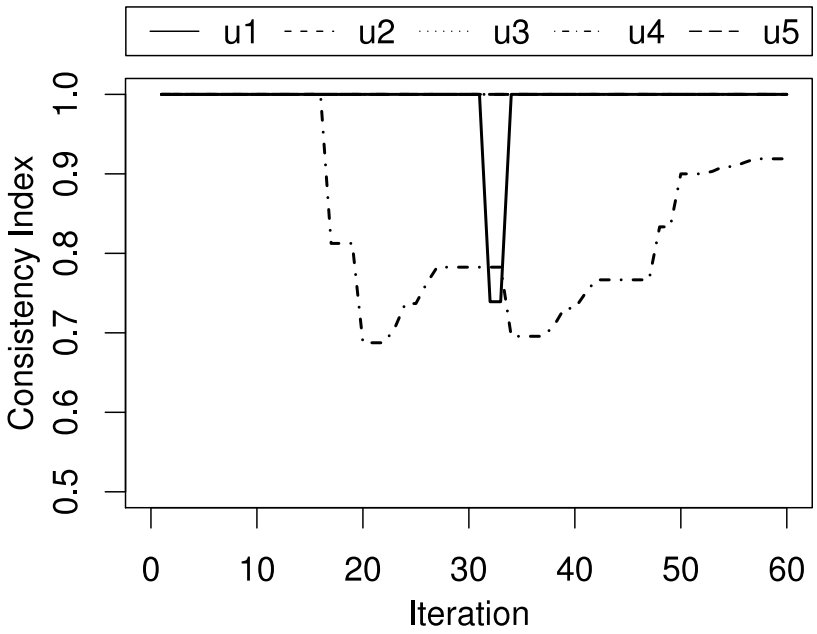


Figura 4.17: Evolució de la consistència (κ) de cinc usuaris ($u_1, \ldots, u_5$) a través del procés evolutiu per la frase "*En quina llengua han parlat tots plegats*" del clúster 1.

d'ajust de pesos ha finalitzat de manera inconsistent $\kappa < 1$. Val a dir que en els ajustos a nivell global de la prova de viabilitat (Alías, 2006) detallades en l'apartat 3.4.7 es varen descartar un 16.6%.

Llavors, tal com s'ha explicat a l'apartat 4.6.2, per a cada prova que finalitza de manera consistent es determina l'última generació informativa mitjançant l'índex λ. La figura 4.15 mostra un exemple del càlcul de l'última iteració informativa per una frase particular de les proves realitzades on s'hi pot observar com cadascun dels 5 usuaris convergeix entre les iteracions 43 i 51. A partir d'aquest punt, l'índex de certesa segueix decreixent fins a al final del procés sense proporcionar nova informació (ja que s'afegeix soroll). Després d'analitzar els experiments duts a terme, l'última iteració informativa s'ubica al voltant de l'iteració 45 (44.5 ± 4.8) en valor mig.

El grau de convergència i correlació dels pesos obtinguts s'avaluen mitjançant els indicadors ρ i τ per tal de validar la metodologia proposada.

A nivell de convergència, s'obté un índex mitjà $\bar{\rho} = 0.62 \pm 0.11$ cosa que indica que no hi ha una millor solució de pesos en termes absoluts per cada dupla <usuari,frase> sinó que més aviat la solució s'ubica en un front de solucions (combinacions de pesos) vàlides que guarden una semblança relativa entre sí ($\rho > 0.5$) sense ser molt similars ($\rho < 0.75$).

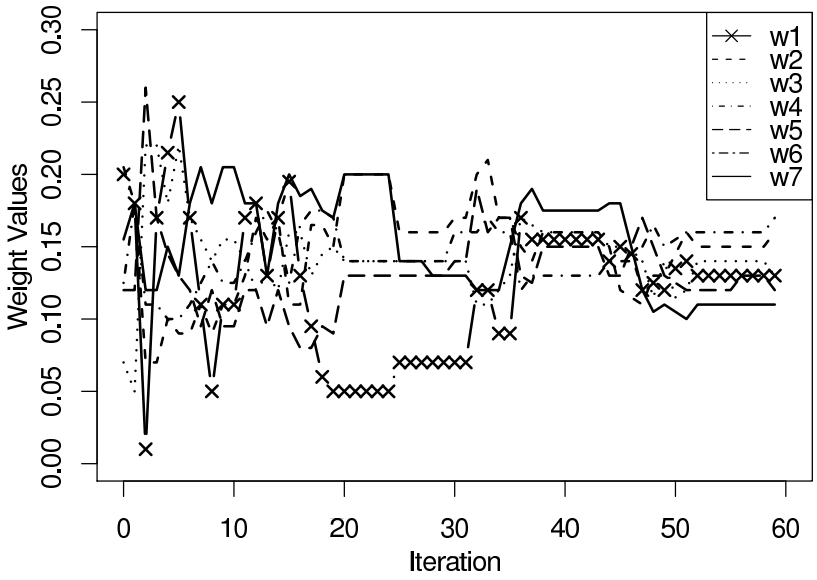


Figura 4.18: Evolució dels valors de pesos normalitzats (mediana entre usuaris) per la frase "*En quina llengua han parlat tots plegats*" del clúster 1. On w_1 representa PIT.T, w_2 ENE.T, w_3 DURL.T, w_4 DURR.T, w_5 PIT.C, w_6 ENE.C i w_7 MFC.C.

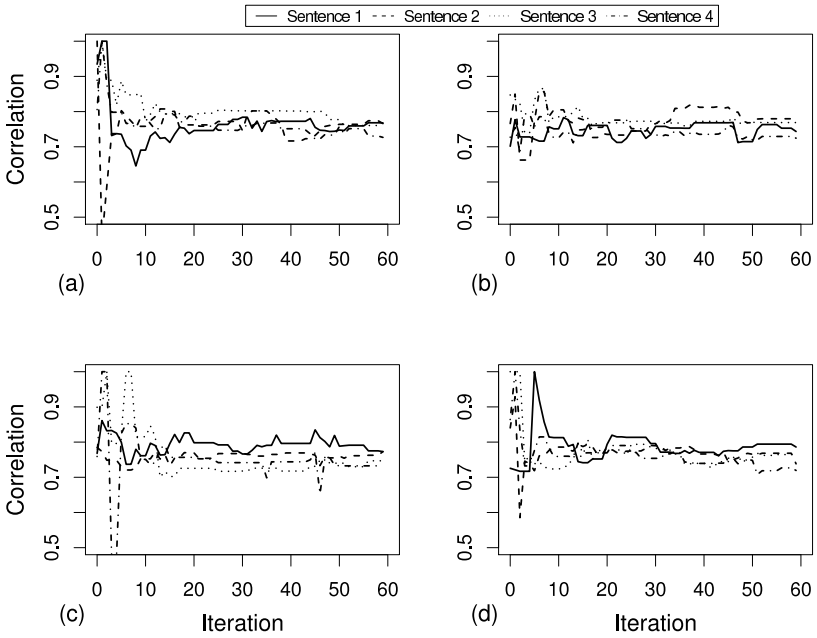


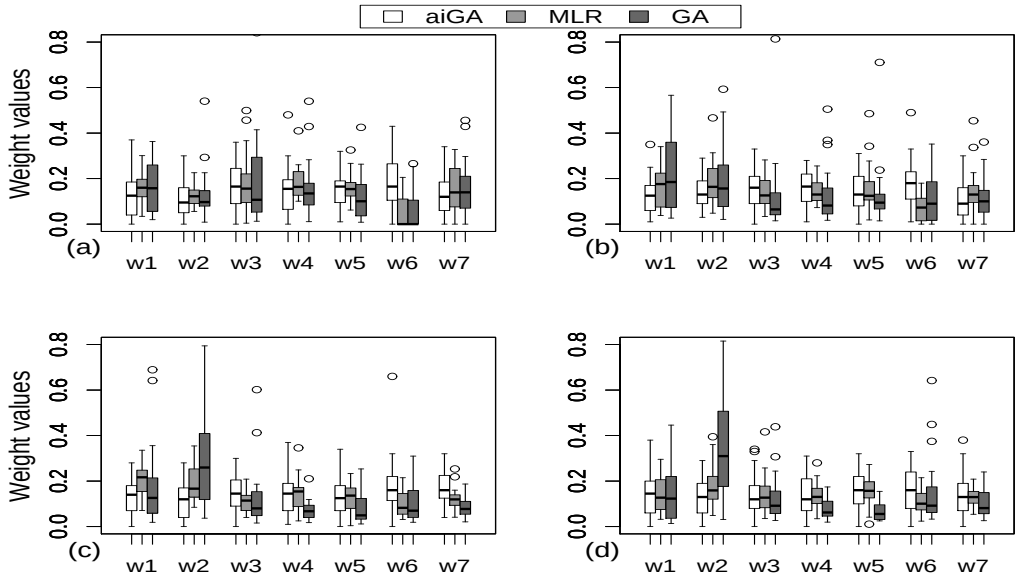
Figura 4.19: Correlació dels valors de pesos (mediana dels diferents usuaris) per les 4 frases seleccionades del (a) clúster 1, (b) clúster 2, (c) clúster 3 i (d) clúster 4.

Addicionalment, malgrat que cada usuari genera un graf particular construït a partir de les seves preferències, la correlació dels millors pesos de cada usuari per a cada clúster també s'avalua, ja que pot donar informació sobre la fiabilitat dels resultats. La figura 4.19 mostra la correlació entre els pesos obtinguts per diferents usuaris detallats per cada una de les 16 frases del test. Cal destacar que, malgrat començar amb valors molt dispersos, els valors de correlació inter-usuari són de $\tau = 0.76 \pm 0.02$ al final del procés iteratiu.

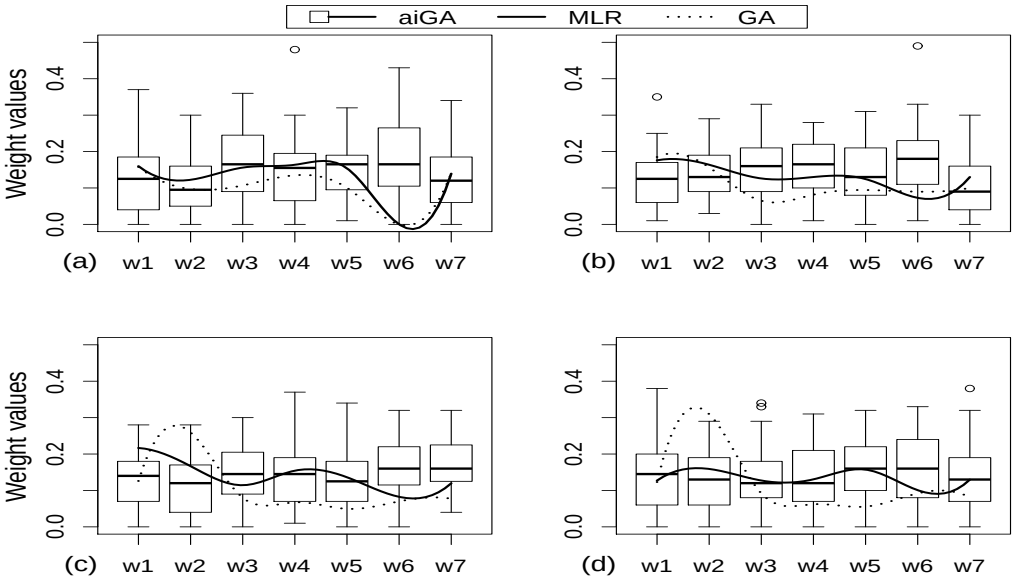
A tall d'exemple, la figura 4.18 mostra els valors de pesos amitjanats en el transcurs del procés per una frase particular. Tal com es pot observar, els valors de pesos comencen amb un patró de pesos sorollós (des de la primera iteració fins a la iteració 20 aproximadament), però posteriorment convergeixen cap a valors més estables, al voltant de la iteració 50. Cal destacar que els pesos de les diferents frases presenten comportaments semblants en les evolucions.

Resultats finals

Finalment, la figura 4.20 mostra els *boxplots* dels patrons de pesos resultants obtinguts per cada clúster. Juntament amb els patrons obtinguts mitjançant aiGA es mostren els patrons de pesos obtinguts mitjançant les tècniques d'ajust automàtic (MLR i GA). Val a dir que aquests patrons de pesos s'obtenen a nivell d'unitat, i després s'agrupen segons les regles de *clustering* obtingudes en l'apartat 4.4.4. En la figura 4.20(b) es pot observar com els mètodes automàtics (GA i MLR) segueixen patrons similars entre ells en les unitats sordes -(a) i (b)- i diferenciats en les unitats sonores. En canvi, en les unitats sonores -(c) i (d)- presenten més similitud els pesos obtinguts amb aiGA i MLR entre ells que els pesos obtinguts amb GA, que accentuaven molt la importància de les durades. Es pot observar tanmateix (figura 4.20(a)) que els pesos obtinguts amb MLR i aiGA adopten patrons amb desviacions estàndard similars ($\sigma_{\text{MLR}} = \pm 0.07$ i $\sigma_{\text{aiGA}} = \pm 0.08$), altrament els pesos obtinguts amb GA presenten uns patrons de pesos més variats degut a la seva cerca elitista en entorns sorollosos ($\sigma_{\text{GA}} = \pm 0.12$). Aquestes dades permeten concloure que el GA realitza una cerca no lineal molt més elitista a nivell local, sense considerar paràmetres supra-segmentals com la concatenació o el ritme i en canvi les cerques mitjançant MLR i aiGA són més suavitzades degut als amitjanats aplicats. En concret MLR amitjana valors degut a la seva naturalesa d'optimització lineal i l'aiGA amitjana valors al considerar clústers d'unitats (i no unitats aïllades) des d'un punt de vista perceptiu, que permet tenir una visió més global (més enllà de la unitat) del comportament dels pesos.



(a) Diferències de patrons considerant les desviacions típiques.



(b) Diferència de patrons de pesos entre les diferents metodologies. Les línies de punts per MLR i GA obeeixen a corbes *spline* ajustades a la mediana de cada pes

Figura 4.20: *Boxplots* dels valors dels pesos obtinguts després d'aplicar els mètodes d'ajust de pesos aiGA, MLR i GA al (a) clúster 1, (b) clúster 2, (c) clúster 3 i (d) clúster 4. Cal tenir en compte que w_1 s'associa a DURL.T, w_2 a DURR.T, w_3 a ENE.C, w_4 a ENE.T, w_5 a MFC.C, w_6 a PIT.C i w_7 a PIT.T.

4.7 Validació de l'ajust perceptiu dels pesos usant aiGA

Per finalitzar l'estudi, es procedeix a validar perceptivament els pesos obtinguts mitjançant aiGA respecte els pesos dels altres tres mètodes de referència (GA, MLR i MOS-*Postmapping*). Primerament, es compara l'aiGA amb els dos altres mètodes objectius (MLR i GA), convé recordar que l'ajust de l'aiGA és a nivell de clúster i l'ajust MLR i GA és a nivell d'unitat. En aquesta prova també es comparen els mètodes automàtics (GA, MLR) entre ells per validar perceptivament quin és millor.

Les dades obtingudes a través d'aquestes comparacions (aiGA vs. MLR, aiGA vs. GA, GA vs. MLR) serveixen per entrenar el tercer mètode perceptiu: MOS-*Postmapping* (apartat 2.4.3), que realitza un ajust perceptiu a nivell global.

Per finalitzar, es comparen els pesos perceptius obtinguts (MOS-*Postmapping*) amb els pesos aiGA. Obtenint així un total de quatre comparatives que avaluen tècniques d"ajust de pesos de dues naturaleses diferents (perceptiva i objectiva). L'avaluació es realitza mitjançant proves de preferència CMOS (Alvarez i Huckvale, 2002; Sityaev *et al.*, 2006).

4.7.1 MOS-*Postmapping*

Es volen comparar els pesos obtinguts amb aiGA amb els pesos obtinguts a través del MOS-*Postmapping* per així avaluar l'idoneïtat de l'aiGA respecte un altre mètode perceptiu. Per tal d'implementar l'esmentada tècnica de referència s'han de recopilar les preferències subjectives dels usuaris segons diferents frases sintètiques generades prèviament de manera *off-line* (els usuaris no cooperen en la cerca, simplement la validen). En aquest cas, aquesta informació s'obté a partir de la prova de preferència (CMOS) realitzada que involucra aiGA, MLR i GA. Conseqüentment, s'obtenen tres valors de MOS per a cada frase (MOS_{aiGA}, MOS_{MLR}, MOS_{GA}) seguint el mateix esquema descrit a (Chu i Peng, 2001; Peng *et al.*, 2002). Tanmateix, els valors MOS no s'obtenen de manera directa sinó que es construeixen a partir de les comparacions CMOS. Aquest fet és degut a que la mida reduïda del corpus (8 min.) no permet realitzar un escombrat d'avaluacions canviant la mida del corpus tal com s'especifica en els treballs referenciats. El procés es realitza segons el procediment següent: *i*) els valors CMOS obtinguts es transformen a un interval absolut amitjanant les puntuacions de l'usuari per a cada dupla <frase,mètode>. Llavors, *ii*) aquests valors es mapen en una escala MOS de 5 punts (de 1 a 5) a través d'una normalització simple *max-min* (equació 4.2).

Seguint el procés descrit a (Chu i Peng, 2001; Peng *et al.*, 2002) es recuperen els fitxers

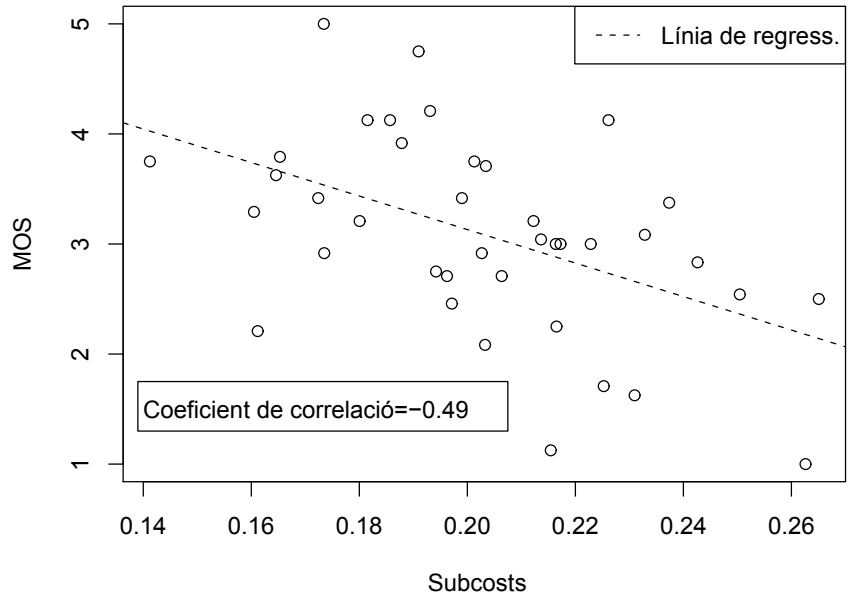


Figura 4.21: Regressió multilineal (MOS-*Postmapping*) entre els subcostos amitjanats i les puntuacions MOS obtingudes a partir de la recopilació de preferències d'usuaris que avaluen diferents frases de test (veure l'apartat 4.7.2).

de log del procés de síntesi per a cada dupla <frase,mètode>. Per a cada subcost s'obté un valor de cost total amitjanant el valor que pren l'esmentat subcost a través de les diferents unitats de la frase. A continuació, el vector de subcostos obtingut es mapa amb les puntuacions MOS de l'usuari emprant un algorisme de regressió multilineal (MLR) (Chu i Peng, 2001; Peng *et al.*, 2002). Tal i com s'ha explicat en l'apartat 2.4.2, pel fet de no permetre valors de pesos negatius, aquesta regressió s'implementa mitjançant l'algorisme NNLS, el qual assegura valors positius de pesos en la regressió. Un cop realitzada l'esmentada associació de subcostos-puntuacions perceptives s'obté una recta de regressió amb una correlació de -0.49 ($R^2 = 0.24$) (veure figura 4.21) - que és un valor similar al que s'obté a (Toda *et al.*, 2006) quan només es consideren els subcostos obtinguts en les millors síntesis (alta puntuació MOS – interval baix de subcostos). Els valors dels pesos obtinguts en l'ajust per MOS-*Postmapping* són els següents: DURL.T $= 0.12$, DURR.T $= 0$, ENE.C $= 0.20$, ENE.T $= 0$, MFC.C $= 0.03$, PIT.C $= 0.65$ i PIT.T $= 0$.

4.7.2 Experiments i resultats

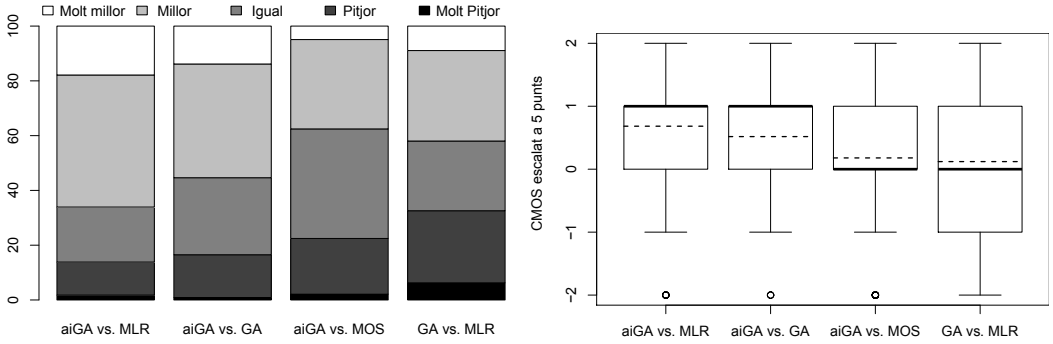
L'objectiu de la validació és avaluar la qualitat de les diferents tècniques, d'ajust de la funció de cost, exposades per un mateix sistema de síntesi de veu. Això permet validar l'impacte de les modificacions incloses en la metodologia d'ajust de pesos mitjançant l'aiGA en comparació a la mateixa comparativa (aiGA vs. GA i MLR vs. iGA) presentada en anteriors treballs (veure apartat 3.4.8), però canviant la metodologia d'ajust iGA per la metodologia MOS-*Postmapping*. A més, es vol comparar l'aiGA amb una altra tècnica d'ajust perceptiu, ja que fins al moment només s'ha comparat l'aiGA amb diferents tècniques d'ajust automàtic (exceptuant la comparació amb iGA), en aquest cas s'ha escollit la tècnica d'ajust MOS-*Postmapping* ja que és la que s'ha treballat més durant el transcurs dels anys (Chu i Peng, 2001; Peng *et al.*, 2002; Toda *et al.*, 2006). A tal efecte, s'han escollit 20 frases diferents del corpus (diferents de les emprades per a realitzar els ajustos perceptius) i s'han realitzat dues proves de preferència: *i*) {MLR vs. GA, aiGA vs. MLR, aiGA vs. GA} que serveixen per obtenir dades per entrenar el mètode de MOS-*Postmapping* i realitzar la segona comparativa *ii*) MOS-*Postmapping* vs. aiGA. Aquestes proves les han realitzat 19 usuaris avaluadors (14 que provenen del grup que va realitzar l'ajust de pesos més 5 nous usuaris que s'inclouen com a grup de control).

En les ambdues proves, se'ls hi demana als avaluadors que comparin dues síntesis candidates per la mateixa frase en termes similars a la comparativa realitzada mitjançant CMOS. No obstant això, a efectes de simplificar el procés d'avaluació, en comptes d'oferir una escala d'avaluació de 7 punts se n'ofereix una de 5 punts que compren l'interval [-2 (clarament pitjor – una respecte l'altra), -1 (pitjor), 0 (igual), 1 (millor), 2 (clarament millor)].

Resultats de la validació perceptiva

La figura 4.22 mostra els resultats obinguts de les comparacions per parelles (mètode A vs. mètode B) per les quatre aproximacions d'ajust de pesos estudiades. A la figura 4.22(a) es pot observar com la proporcionalitat de millor o molt millor de l'aiGA guanya a la proporcionalitat de pitjor o molt pitjor en la majoria de comparacions – el terme millor de la figura fa referència a la tècnica de l'esquerra en la comparació (p.ex. en la comparació aiGA vs. MLR millor vol dir aiGA $>$ MLR). La figura 4.22(b) mostra els resultats en forma de *boxplots*. En aquest cas, els resultats mostren una puntuació positiva (> 0) quan es selecciona la tècnica de l'esquerra i una puntuació negativa (< 0) quan es selecciona la tècnica de la dreta.

En els resultats es pot observar com els usuaris prefereixen clarament els pesos basats



(a) *Barres de preferència.* (b) *Boxplots de preferència.*

Figura 4.22: Resultats del CMOS escalat segons cinc punts que recull les preferències dels usuaris quan es comparen frases sintetitzades amb les configuracions de pesos aiGA contra les configuracions de pesos MLR, GA i MOS-*Postmapping* (indicat com a MOS). A més, s'afegeix com a referència la comparativa GA vs. MLR. La línia de punts horitzontal dins dels *boxplots* representa els valors mitjans de les distribucions.

amb aiGA respecte els pesos obtinguts amb mètodes d'ajust automàtics (MLR i GA amb distàncies cepstrals). A més, hi ha certa tendència a preferir l'aproximació mitjançant aiGA respecte l'altra tècnica perceptiva, MOS-*Postmapping*. Concretament, hi ha un 37.54 % de preferència pels pesos obtinguts mitjançant aiGA respecte un 22.46% de preferència pels pesos obtinguts mitjançant MOS-*Postmapping*. Tanmateix, la comparativa dels dos mètodes perceptius tendeix a generar una qualitat sintètica similar (40% de comparacions iguals en la figura 4.22(a)) .

Addicionalment, a efectes d'avaluar la validesa estadística d'aquest resultats, s'ha realitzat una anàlisi mitjançant una prova t de Student per parelles, la qual estableix la significança de les preferències en cada parella de mètodes d'ajust. Com a resultat, les proves mostren que aiGA $>$ MLR (mediana$=1$ i mitjana$=0.68$) amb un nivell de confiança de $p < 2 \cdot 10^{-16}$, aiGA $>$ GA amb un nivell de confiança de $p = 8.7 \cdot 10^{-13}$ (mediana$=1$ i mitjana$=0.51$). Alhora, aiGA $>$ MOS-*Postmapping* és també significativa amb un nivell de confiança de $p = 0.00083$ (mediana$=0$ i mitjana$=0.18$). Finalment, es pot observar com la diferència entre MLR i GA no es estadísticament significativa, ja que $p > 0.05$ amb mediana$=0$ i mitjana$=0.16$.

En definitiva, aquests resultats reforcen la conclusió que l'aiGA millora en termes de qualitat sintètica resultant els mètodes d'ajust de pesos automàtics. Tanmateix, respecte l'altre mètode d'ajust perceptiu, es pot observar com l'aiGA té certa tendència a millorar

el MOS-*Postmapping* però que en termes generals ambdós mètodes presenten una qualitat similar. En últim terme, però no menys important, cal afegir que el grup de control format pels 5 usuaris que no van formar part de l'ajust mitjançant aiGA presenten un comportament similar: aiGA $>$ MLR amb un nivell de confiança de $p = 1.9 \cdot 10^{-4}$, aiGA $>$ GA MLR amb un nivell de confiança de $p = 2 \cdot 10^{-3}$, aiGA $>$ MOS-*Postmapping* MLR amb un nivell de confiança de $p = 0.036$. A més, no s'observen diferències significatives entre MLR i GA. Per tant, es pot concloure que els patrons de pesos obtinguts mitjançant aiGA porten a frases amb una qualitat sintètica més elevada que els mètodes de referència fins i tot apreciada pels avaluadors que no han participat en l'ajust de pesos usant l'aiGA. És a dir, els pesos obtinguts pels diferents usuaris durant l'ajust, són bons per usuaris que no han participat en l'esmentat ajust.

Comparació amb els resultats de la prova de viabilitat a nivell global

Els resultats que s'han presentat fins ara presenten una comparativa dels diferents mètodes d'ajust de pesos en termes absoluts. Malauradament, no es poden comparar de manera directa la prova de viabilitat de l'estat de l'art (Alías, 2006) explicats en l'apartat 3.4.8 ja que els resultat anteriors segueixen una metodologia diferent d'anàlisi. A part, convé recordar que Alías (2006) no inclou el MOS-*Postmapping*. Cal esmentar que els resultats de la prova de viabilitat no contenien estudis de significança estadística. No obstant això els resultats eren similars independentment de la frase avaluada (5 locucions) cosa que dotava els resultats de certa robustesa. També cal afegir que el nombre d'usuaris avaluadors (10) era menor que l'emprat en l'estudi actual (19). Els resultats obtinguts a (Alías, 2006) es presenten en la figura 4.23.

A efectes de comparar els nous resultats amb els de la prova de viabilitat, per cada dupla <usuari,frase> es considera només la metodologia d'ajust de pesos guanyadora en termes absoluts (millor puntuació CMOS) tal i com es va fer a (Alías, 2006). Així es poden comparar els resultats de la mateixa manera. Els resultats adaptats es mostren a la figura 4.24.

En la comparativa entre els resultats dels dos estudis es pot observar com, en ambdues avaluacions, l'aiGA resulta la metodologia d'ajust de pesos guanyadora. Tanmateix, en l'estudi realitzat per (Alías, 2006) es pot veure com la metodologia MLR era la rival principal de l'aiGA, superant a l'iGA i el GA, mentre que en el nou estudi queda relegada en últim terme essent superada pel GA (tot i que la figura 4.22(b) es veu que aquesta diferència no és significativa). En canvi, es pot observar com la metodologia d'ajust de pesos perceptiu basada en MOS-*Postmapping* supera les metodologies automàtiques basades en

distàncies cepstrals però no arriba al nivell d'acceptació de l'aiGA. En termes generals cal destacar que en les noves proves els resultats no són tant uniformes ja que els resultats depenen en gran mesura de la frase avaluada. En aquest sentit val a dir que la prova de viabilitat es va realitzar amb un nombre reduït de frases (5 locucions).

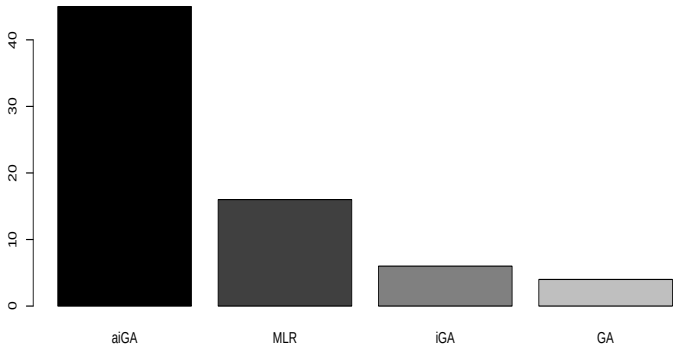
4.8 Conclusions

En aquest apartat s'argumenta l'impacte de les diferents contribucions plantejades durant el transcurs d'aquest capítol així com la viabilitat de la metodologia d'ajust de pesos proposada respecte les altres metodologies objectives o perceptives considerades.

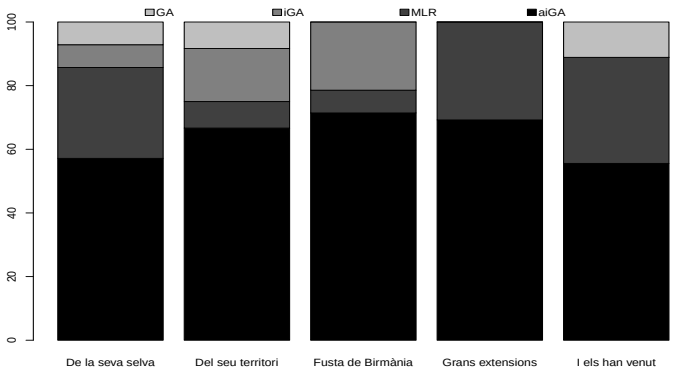
Primerament, a l'apartat 4.2, s'ha realitzat un estudi estadístic del corpus tant en termes prosòdics com en termes de subcostos acústics associats a la funció de cost plantejada en l'apartat 2.2.3. En aquest apartat, s'ha detectat la pobra normalitat de les distribucions dels subcostos provocant, a priori, una mala predisposició a ser modelades per mètodes de regressió lineal clàssics com és el cas del MLR, i en conseqüència el MOS-*Postmapping*. Per tal de minimitzar el problema, s'ha introduït un canvi de la funció de normalització de subcostos per tal de millorar lleugerament la normalitat d'aquestes dades malgrat que no s'ha pogut assolir una normalitat plena de les dades.

Un cop analitzades les dades amb les que treballa la funció de cost, s'ha analitzat la fiabilitat dels pesos (veure apartat 4.3) obtinguts amb les metodologies clàssiques d'ajust de pesos basades en distàncies cepstrals (MLR i GA). Segons els estudis realitzats es pot observar que els pesos obtinguts amb la tècnica d'ajust MLR presenten un baix coeficient de determinació R^2, cosa que indica la baixa fiabilitat del model degut al comportament altament sorollós que presenten els diferents subcostos a través de les diferentes versions de les unitats enregistrades. En canvi, s'ha pogut observar com el GA és capaç d'inferir tendències en l'ajust de pesos malgrat la seva alta sorollositat. Donat que no es coneix cap índex de robustesa de l'ajust anàleg al coeficient de determinació que s'empra amb MLR s'ha analitzat l'evolució del *fitness* en el transcurs de les generacions juntament amb la desviació típica que pren el valor del pes en el transcurs de 2000 iteracions. Aquesta anàlisi de la desviació típica es pot dur a terme ja que quan el nombre d'iteracions tendeix a infinit el seu valor oscil·la sobre el mateix valor mig central (Goldberg, 1989). Els resultats de l'estudi han mostrat una millor fiabilitat dels pesos que s'obtenen mitjançant GA, gràcies a que el mètode és capaç de trencar la linealitat de les dades que imposa el MLR, tal i com s'havia esbossat en l'apartat 3.2.4 (Alías i Llorà, 2003).

Posteriorment, a l'apartat 4.4 s'ha introduït la problemàtica del nivell d'ajust de pe-

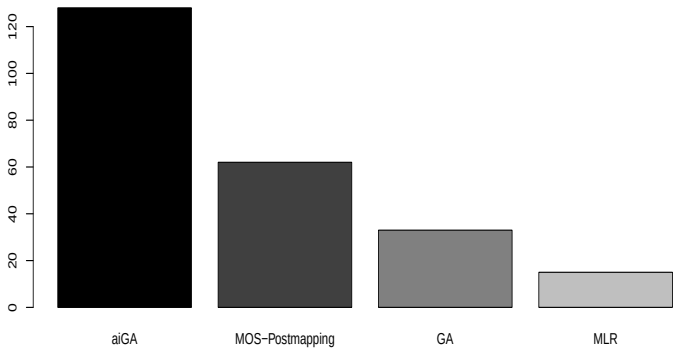


(a) Distribució dels mètodes d'ajust guanyadors per usuari de la validació perceptiva.

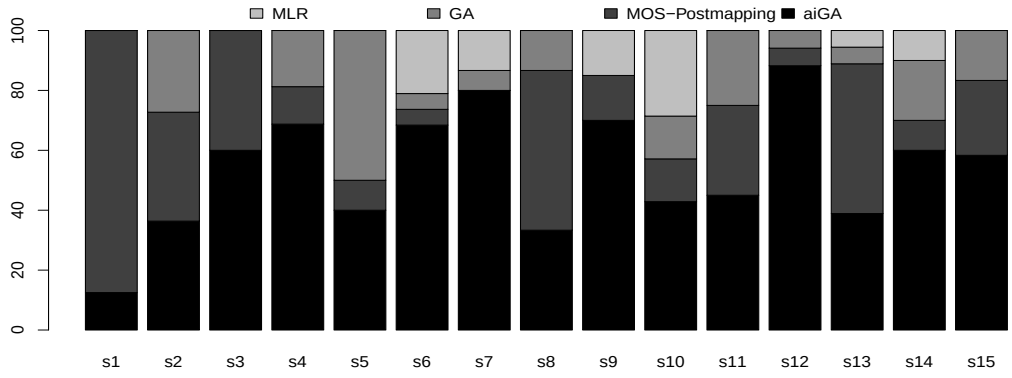


(b) Comparativa dels diferents mètodes detallats per frase.

Figura 4.23: Comparativa perceptiva entre els diferents mètodes d'ajust de pesos segons (Alías, 2006).



(a) Distribució dels mètodes d'ajust guanyadors per usuari de la validació perceptiva.



(b) Distribució dels mètodes d'ajust guanyadors per usuari de la validació perceptiva detallat per a totes les frases.

Figura 4.24: Comparativa perceptiva entre els diferents mètodes d'ajust de pesos segons les contribucions descrites en aquest capítol (normalització sigmoide lineal i ajust a nivell de clúster).

sos, considerant que els mètodes clàssics d'ajust automàtic respecten l'especificitat de cada unitat a l'hora de ponderar els subcostos (nivell d'unitat), mentre que el mètodes d'ajust perceptiu generalitzen només un vector de pesos per a totes les unitats analitzades (nivell global). En aquest punt es presenta un nivell d'ajust intermedi basat en *clustering* mitjançant arbres de decisió. Aquest nou enfoc té en compte els patrons de pesos obtinguts de manera automàtica mitjançant GA. L'estudi també analitza el número i mida òptims dels clústers emprant índexs típics que avaluen la validesa del *clustering* (Günter i Bunke, 2003). En aquest sentit s'ha observat que no resulta fàcil fixar el nombre de clústers ja que cadascuna de les mètriques proposa un nombre de clústers diferent. No obstant això, la majoria de mètriques consideren un nombre de clústers que oscil·la entre 2 i 4.

Per tal de desacoblar els efectes del post-processament del senyal en la composició de la forma d'ona (veure apartat 2.2.4), en l'apartat 4.4.3 s'analitza com el processament del senyal de veu pot emmascarar l'optimització de la selecció d'unitats. Per analitzar aquest impacte es compara en l'ajust interactiu els ajustos de pesos la tècnica clàssica de TD-PSOLA amb una tècnica simple de concatenació de les unitats amb un mínim ajust de fase (coneguda com a WAV) on posteriorment (apartat 4.6) es confirma que realitzar l'ajust de pesos amb processament del senyal introdueix ambigüitat (emmascara el criteri) en l'ajust.

En l'apartat 4.5 es justifica la necessitat d'ampliar els indicadors de fiabilitat de l'aiGA més enllà del seu indicador de consistència κ definit en l'apartat 3.4.7. A tal efecte, es defineixen nous indicadors per tal de mesurar l'idoneïtat dels pesos obtinguts (Formiga *et al.*, 2010). Aquests indicadors analitzen diversos efectes en l'evolució interactiva, tals com la certesa (indicador λ), el grau de convergència de l'usuari (indicador ρ) o la correlació dels pesos obtinguts per diferents usuaris (indicador τ). Les proves realitzades entre WAV i TD-PSOLA demostren que l'indicador de consistència κ resultava insuficient per avaluar la bondat de les solucions obtingudes. No obstant això, mitjançant els indicadors a nivell de genotip (ρ i τ) no s'aprecien diferències degut a la introducció de l'ambigüitat, demostrant la validesa de l'aiGA per optimitzar entorns sorollosos o ambigus.

Posteriorment, en l'apartat 4.6, un cop identificats els clústers de pesos que adopten patrons similars, s'optimitzen perceptiblement els pesos mitjançant aiGA. Com ja s'ha dit, segons la mètrica λ es conclou que els pesos obtinguts sense postprocessament del senyal (WAV) presenten un millor índex de certesa respecte els que si incorporen el postprocessament (TD-PSOLA). Posteriorment s'analitzen els pesos obtinguts per cada clúster i es conclou que el GA realitza una cerca no lineal i elitista a nivell local mentre que el MLR i l'aiGA obtenen valors més suavitzats. Aquest fet ve donat perquè MLR amitjana valors degut a la seva naturalesa d'optimització lineal i l'aiGA amitjana valors al considerar

clústers d'unitats (i no unitats aïllades) des d'un punt de vista perceptiu, que permet tenir una visió més global (més enllà de la unitat) del comportament dels pesos.

En últim terme (a l'apartat 4.7), es realitza una validació de l'ajust mitjançant aiGA en tres etapes: *i*) es confronten els pesos obtinguts mitjançant aiGA amb els pesos obtinguts mitjançant GA i MLR per tenir una primera validació del mètode proposat respecte dos mètodes d'ajust objectiu. En les proves es confirma la superioritat de l'aiGA respecte GA i MLR. Un cop validat l'ajust mitjançant aiGA respecte les tècniques automàtiques, *ii*) es procedeix a obtenir pesos mapant linealment els subcostos amb les puntuacions perceptives obtingudes. Aquesta tècnica, implementada mitjançant l'algorisme NNLS (veure apartat 2.4.2), es coneix com a MOS-*Postmapping* (Chu i Peng, 2001; Peng *et al.*, 2002) i serveix com a base de comparació perceptiva respecte l'aiGA. També es pot considerar el MOS-*Postmapping* com una evolució de l'ajust de pesos automàtic basat en MLR, però substituint la distància cepstral objectiva per un conjunt de puntuacions MOS obtingudes en una prova d'avaluació amb usuaris reals. Al final es realitza una segona validació *iii*) que comparara l'aiGA amb el MOS-*Postmapping*. En aquesta última comparació l'aiGA adopta una lleugera millora estadísticament significativa respecte el MOS-*Postmapping* tot i que en la comparació entre ambdós mètodes predomina la igualtat.

En resum s'han presentat contribucions per tal de millorar l'eficiència i fiabilitat de l'ajust de pesos per la funció de cost a nivell de: *i*) normalització estadística, *ii*) estudi de la fiabilitat en els modelats basats en distàncies cepstrals, *iii*) nivell d'ajust de pesos, *iv*) postprocessament del senyal en l'ajust i *v*) indicadors per estudiar la fiabilitat dels processos evolutius interactius. L'objectiu principal s'ha mantingut en tot moment: veure la viabilitat i fiabilitat de la metodologia proposada (GA+CART+aiGA). La fiabilitat dels resultats s'ha avaluat mitjançant diversos indicadors (R^2, σ, κ, λ, ρ i τ) i la viabilitat dels resultats s'ha estudiat fent una comparativa CMOS amb altres tècniques d'ajust de pesos: MLR i GA d'ordre objectiu i MOS-*Postmapping* d'ordre subjectiu. A més, s'ha introduït un procés de *clustering* que permet l'agrupació de les unitats que participen en l'ajust subjectiu de pesos en clústers que presenten un patró de comportament de pesos similar. Així es pot realitzar un ajust perceptiu dels pesos respectant la diversitat del corpus d'una manera assolible. Al final es pot observar, que segons proves de preferència, els pesos obtinguts de manera perceptiva a través d'aiGA tenen una millor acceptació que els pesos obtinguts mitjançant les altres tècniques, bé siguin aquestes objectives o perceptives.

El punt de partida del MOS-*Postmapping* és mapar de manera lineal la distància objectiva amb la percepció humana. Malgrat ser una aproximació interessant, la metodologia basada en aiGA millora alguns aspectes d'aquesta simple linealització. En primer

lloc, el MOS-*Postmapping* només permet inferir una regressió amb restriccions lineals (o polinòmiques) entre subcostos i les proves percepció fent-la sensible al soroll. Per tant, qualsevol relació no lineal (p.ex. sorollosa) o correlació dels subcostos amb la percepció humana es pot assolir amb un enfocament basat en aiGA. En segon lloc, malgrat que el número de frases emprades en el test MOS subjectiu que necessita la tècnica de MOS-*Postmapping* es pot augmentar fàcilment, l'estat de l'art indica que, per tal d'aconseguir una optimització fiable, les locucions utilitzades per a l'experiment de MOS s'haurien de dissenyar de manera específica per tal que les unitats considerades cobrissin àmpliament la variabilitat prosòdica per tots els subcostos (Peng *et al.*, 2002; Toda *et al.*, 2003), per a evitar obtenir (Toda *et al.*, 2003) combinacions de pesos il·lògiques o excessivament esbiaixades a les locucions d'entrenament (Toda *et al.*, 2006).

Finalment, es conclou que els indicadors amb els que s'ha treballat en aquest capítol (κ, λ, ρ, τ) permeten obtenir informació sobre consistència, certesa (o ambigüitat), convergència i correlació dels pesos obtinguts a través del procés evolutiu. Gràcies a aquests indicadors es disposa d'una metodologia implícita que permet millorar l'eficiència de l'ajust de pesos, assegurant en tot moment la consistència d'usuari. En aquest sentit, gràcies a aquests indicadors es pot concloure que incorporar el mòdul de postprocessament del senyal (TD-PSOLA) després de realitzar la selecció, introdueix ambigüitat en l'ajust ja que emmascara les diferències que puguin percebre els usuaris.

Per finalitzar el treball d'aquest capítol a la figura 4.25 es mostra la metodologia d'ajust proposada denotant entre parèntesi la tècnica concreta que s'utilitza en cada pas.

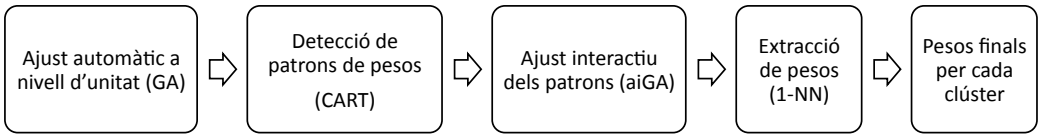


Figura 4.25: Diagrama de les etapes de la metodologia proposada per ajustar pesos a nivell de clúster. Entre parèntesi es destaca la tècnica proposada en cada etapa.

Per finalitzar, es fa notar que gran part del treball realitzat en aquest capítol ha estat publicat recentment a Alías *et al.* (2011).

4.9 Aspectes pendents

Aquest capítol ofereix unes primeres contribucions per tal de millorar l'ajust de la funció de cost de la selecció d'unitats de manera perceptiva. Aquestes contribucions es sostenen en tot moment sobre la base empírica que l'entorn d'optimització perceptiva és un entorn sorollós (un usuari es pot contradir a sí mateix a l'hora d'avaluar la qualitat d'una síntesi o diversos usuaris poden entrar en diferents contradiccions). Malgrat aquestes contribucions inicials que aporten uns resultats prometedors, s'ha de reconèixer que quan la metodologia d'ajust proposada (aiGA) entra a competir amb altres metodologies d'ajust perceptiu de l'estat de l'art, la millora no és espectacular sinó que més aviat es tendeix a una lleugera tendència de millora. Per tant, cal identificar certs problemes que no han estat resolts en aquest primer bloc de contribucions per tal de ser resolts més endavant. Els diferents problemes detectats s'exposen a continuació:

i) En aquest capítol es presenta una modificació de la normalització (i transformació) dels subcostos per tal de millorar la seva gaussianitat. Tanmateix, aquesta modificació no permet assolir una plena normalitat de les dades deixant obert per a un treball futur millorar la transformació amb l'objectiu de millorar els índexs de normalitat dels subcostos.

ii) A l'hora d'analitzar la fiabilitat obtinguda pels mètodes objectius d'ajust de pesos segons les distàncies cepstral s'obtenen índexs de fiabilitat bastant baixos. Concretament MLR obté un coeficient de determinació mitjà de $R^2 = 4.6\%$ i GA presenta un nivell de soroll aproximat sobre el valor mig del 25%. Degut a aquesta falta d'eficiència en el modelat de les dades, s'obre la pregunta sobre si l'ajust de pesos a nivell d'unitat resulta suficient ja que no respecta l'especificitat que pren la mateixa unitat en diferents contextos. Tal com s'ha exposat en l'apartat 2.3.4 la forma d'ona que conforma una unitat pot variar molt depenent de la seva ubicació contextual (veïnatge fonètic, parametrització lingüística, etc.).

iii) En aquest capítol només s'han analitzat 7 subcostos de la funció de cost de naturalesa acústica i contínua: 4 de *target* i 3 de concatenació (de tipologia ASF en l'apartat 2.3.4). Per tant, queda obert l'estudi de l'aplicabilitat de l'ajust de pesos mitjançant aiGA mesclant els subcostos emprats (ASF) amb subcostos lingüístics de naturalesa discreta (de tipologia IFF en l'apartat 2.3.4). A més, s'ha d'afegir el fet que la prova de viabilitat realitzada usa un corpus petit (8 min.) que no es pot considerar en sí mateix com un corpus apte per la selecció d'unitats, per tant queda pendent l'estudi de la resposta de

la metodologia d'ajust de pesos basada en aiGA aplicat en l'ajust de pesos d'un corpus de mida més acceptable dins l'àmbit de la selecció d'unitats ($\geq$ 1 h.).

iv) En l'estudi realitzat es pot observar també com els valors de pesos obtinguts segons les diferents metodologies difereixen entre sí. Cal tenir en compte que quan s'obtenen els patrons de pesos inicials segons les distàncies cepstrals es realitza l'assumpció implícita que els pesos obtinguts no guarden relació amb la tècnica de composició de forma d'ona emprada (WAV, TD-PSOLA,...). De tota manera no hi ha una solució per obtenir patrons de pesos perceptius abans de les proves d'ajust perceptives pròpiament dites (esdevenint un cercle viciós). Tanmateix queda oberta la consideració de reconstruir l'arbre de decisió, i per tant reclusteritzar les unitats, un cop realitzades les proves d'ajust perceptiu.

v) De tota manera, malgrat els mètodes d'ajust automàtic no ofereixen de moment una bona inicialització dels pesos perceptius, encara queda camí per recórrer (contribucions de millora exposades en els punts *ii* i *iii*). Si un cop millorades les tècniques d'ajust objectiu (millor normalització, nivell contextual de la unitat) es posen a competir entre elles a nivell perceptiu, es pot obtenir una primera discriminació prèvia de caire perceptiu per escollir quina és la millor metodologia d'ajust objectiu de pesos que cal emprar per tal de realitzar l'agrupament de les diferents unitats abans d'iniciar l'ajust interactiu mitjançant aiGA.

vi) Malgrat el CART és una tècnica de *clustering* coneguda i àmpliament acceptada dins de la comunitat de síntesi de la parla basada en selecció d'unitats (Black i Taylor, 1997*a*), manca per estudiar la seva bondat respecte altres metodologies de *clustering* (*k-means*, *Expectation-Maximization*, etc.) per tal de refinar l'agrupament de pesos. No obstant això, les tècniques alternatives d'agrupació de dades no realitzen l'agrupament i classificació (associació segons l'especificitat fonètica) en una sola etapa. En aquest sentit, es podria considerar realitzar l'agrupament en dues etapes: la primera mitjançant *clustering* pròpiament dit i la segona una assignació dels clústers a l'especificitat fonètica mitjançant CART..

vii) Per últim, l'extracció de resultats un cop realitzades les proves d'ajust (consens) s'ha limitat a una simple recol·lecció del valor mig per cada pes associat a cada subcost. En un estudi més complet, quedaria pendent estudiar una extracció de resultats basada també en l'agrupament de criteris dels usuaris avaluadors que considerés algun tipus de model més enllà de la mitjana.

CAPÍTOL 5

Escenari real: corpus mitjà amb subcostos acústics i lingüístics

5.1 Introducció

La revisió de la prova de viabilitat (Alías, 2006) d'aplicar aiGA per l'ajust dels pesos de la funció de cost de sistemes CTP-SU ha comportat les primeres contribucions (preliminars) d'aquesta tesi doctoral.

Primerament s'ha proposat una millora de la normalització dels subcostos a la funció de cost. Posteriorment, s'ha proposat realitzar l'ajust perceptiu a nivell de grup o clúster per respectar el comportament diferent dels subcostos en funció de l'especificitat fonètica de la pròpia unitat. La tercera contribució ha estat la proposta de nous indicadors per mesurar aspectes de les proves evolutives més enllà de la consistència dels usuaris (ambigüitat, convergència, correlació, etc.). La última contribució ha estat la proposta d'una metodologia per extreure un vector de pesos de consens d'entre els millors pesos dels diferents usuaris que realitzen la prova (basada en 1-*Nearest Neigbour*). Addicionalment, s'ha realitzat una comparativa de la qualitat dels pesos obtinguts amb aiGA respecte altres tècniques d'ajust de pesos típiques de l'estat de l'art: MLR/NNLS, GA i MOS-*Postmapping*.

No obstant això, un cop realitzada la revisió, és el moment d'estudiar l'aplicabilitat d'aquesta tècnica en un entorn real de selecció d'unitats així com millorar les contribucions preliminars plantejades en el capítol anterior, plantejant-ne de noves si s'escau.

En aquest sentit, per tal de proporcionar un entorn real de selecció d'unitats i alhora solucionar els aspectes no resolts esmentats en el capítol anterior (apartat 4.9) es formulen una sèrie d'objectius que s'intentaran superar durant el transcurs d'aquest capítol. Els objectius s'exposen a continuació:

i) Proporcionar a l'aiGA un entorn d'optimització real en selecció d'unitats que consideri subcostos de naturalesa diferent (lingüística) i alhora un corpus extens ($>$ 1h – (Taylor, 2009)) i etiquetat automàticament.

ii) Millorar la fiabilitat dels mètodes automàtics (MLR/NNLS, GA) en l'ajust de pesos mitjançant distàncies cepstrals per obtenir millors patrons de pesos a nivell de clúster.

iii) Estudiar l'impacte del context de la unitat en l'ajust dels pesos. Aquest impacte s'estudia tant en els mètodes automàtics com en els mètodes perceptius.

iv) Realitzar un estudi de les diferents metodologies de *clustering* típiques de l'estat de l'art per tal de millorar la detecció automàtica de patrons de pesos.

v) Obtenir un model robust de consens que mapi les preferències dels diferents usuaris davant dels pesos escollits.

vi) Estudiar la vigència dels patrons de pesos obtinguts mitjançant distàncies cepstrals un cop es disposa de pesos obtinguts de manera perceptiva.

vii) Estudiar l'impacte de canviar la funció de cost (distància de Manhattan ponderada o *average* – AVG) a una distància euclídea (també anomenada *Root-Mean Square* – RMS) tal com recomanen els estudis realitzats per Toda *et al.* (2006).

A l'apartat 5.2 es descriu amb detall el corpus de veu emprat en aquest capítol i es justifica la seva elecció. Addicionalment es tracta el problema de la normalitat dels diferents subcostos, proposant millorar la funció de normalització. A l'apartat 5.3 s'introdueixen els subcostos lingüístics, explicant la seva motivació i la seva adequació a un escenari real d'ajust d'unitats (veure apartat 2.3.4), combinant subcostos de diferent naturalesa (acústica, lingüística).

A l'apartat 5.4 s'estudia l'impacte dels diferents contextos fonètics i lingüístics en l'ajust dels pesos de la funció de cost, fent evident la necessitat d'ajustar un vector de pesos a cada context. A tal efecte es proposa un nou nivell d'ajust basat en subunitats contextualitzades. Al mateix apartat s'analitza la millora que aporta el nou nivell d'ajust de pesos respecte la fiabilitat dels pesos obtinguts mitjançant mètodes automàtics d'ajust basats en

distàncies cepstrals. A l'apartat 5.5 s'exposen les limitacions que té la detecció de patrons de pesos mitjançant CART i es proposa separar el procés de detecció de patrons en una primera etapa de *clustering*, i en una segona etapa de mapatge que associa els patrons prèviament obtinguts als diferents contextos lingüístics i fonètics presents en el corpus. En aquest apartat s'inclou un estudi sobre diferents algorismes de *clustering* en funció de les mètriques de bondat de les particions.

A l'apartat 5.6 es proposa emprar models latents per consensuar els criteris dels diferents usuaris de manera robusta degut al seu èxit en modelar comportaments en ciències socials (Gibson, 1959). A l'apartat 5.7 s'explica com es realitza l'ajust perceptiu de pesos mitjançant aiGA monitoritzant el procés evolutiu amb les mètriques explicades en el capítol anterior (capítol 4).

L'esquema nou proposat en aquest capítol es mostra a la figura 5.1.

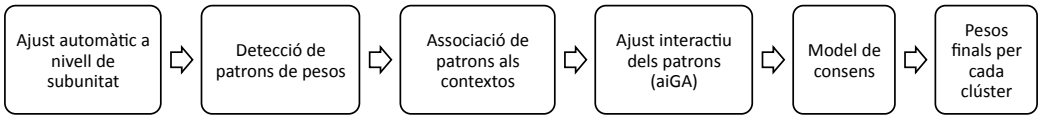


Figura 5.1: Diagrama de les etapes de la metodologia proposada per ajustar pesos a nivell de clúster.

Finalment, a l'apartat 5.8 es descriuen les proves perceptives per validar les contribucions descrites en aquest capítol. En les proves perceptives és on s'estudia l'impacte de canviar la mètrica de la funció de cost (distància de Manhattan ponderada) per una distància euclídea (RMSE — Toda *et al.* (2006)). Un cop estudiades les contribucions i les mètriques es valida l'eficiència de l'aiGA en el problema de l'ajust de pesos de la funció de cost respecte altres mètodes d'ajust de pesos de l'estat de l'art. Concretament es compara amb un mètode objectiu (MLR/NNLS) i amb un mètode perceptiu (MOS-*Postmapping*). A més, per analitzar l'impacte de treballar a nivell de clúster s'introdueixen en la comparativa els pesos ajustats amb aiGA a nivell global: un sol vector de pesos perceptius per a tot el corpus. A més, la introducció d'aquesta comparació permet validar la metodologia aiGA respecte el MOS-*Postmapping* en igualtat de condicions (mateix nivell d'ajust global) a la vegada que proporciona informació sobre la idoneïtat d'escollir els pesos en funció del context de la unitat en el text d'entrada. Al final del capítol, a l'apartat 5.9, es presenten les conclusions.

5.2 Descripció del corpus

Un dels elements clau d'aquest capítol és el corpus de veu. Es deixa de banda el corpus de 8 minuts que ha servit per a demostrar la viabilitat de la proposta des d'un punt de vista de prova de concepte (Alías, 2006) i es passa a usar un corpus d'extensió suficient per obtenir una bona cobertura de diferents unitats a seleccionar. Taylor (2009) diu que la extensió mínima d'un corpus per a que no hi hagin unitats dèbilment poblades (poques versions) ha de ser d'una hora.

En aquest sentit, en aquest capítol es canvia el corpus de treball amb l'objectiu de treballar amb una base de dades més extensa. Primerament s'han emprat corpus del grup de recerca aptes per a la selecció d'unitats (e.g. *url_pat_es* (Iriondo *et al.*, 2008)) a efectes de dissenyar prototipus preliminars (Formiga *et al.*, 2010). Tanmateix, en el treball que conforma aquest capítol de la tesi s'ha optat per treballar amb el corpus ofert als participants de l'última competició de sistemes de conversió de text en parla (CTP) a nivell ibèric (Albayzín - FALA2010 (Méndez Pazó *et al.*, 2010)).

L'elecció d'aquest corpus, en castellà, permet analitzar la viabilitat de la metodologia d'ajust de pesos proposada en un CTP basat en un corpus apte per la selecció d'unitats i etiquetat automàticament (*url_fer_ct* i *url_pat_es* són revisats a mà). A més, aquesta elecció permet comparar resultats amb els altres sistemes que conformen l'estat de l'art en llengua castellana (Méndez Pazó *et al.*, 2010). Aquest corpus, anomenat *uvig_dav_es*, està dissenyat i enregistrat per l'Universitat de Vigo i té una durada d'aproximadament dues hores (1.9h) d'un locutor masculí (David). Val a dir que el corpus s'ha dissenyat expressament per un CTP basat en selecció d'unitats (CTP-SU). El corpus es composa de 1217 locucions formades per 17797 paraules, les quals proporcionen una cobertura per a un vocabulari de 5465 paraules diferents.

Malgrat el fet que el corpus disponible s'ofereixi etiquetat, el grup de recerca va decidir re-etiquetar de nou el corpus amb les seves eines pel fet de no disposar del mateix mòdul de transcripció de grafema-al·lòfon emprat en origen. Aquesta nova parametrització comporta, entre d'altres: *i*) una nova segmentació (alineació de les marques de principi i fi de l'al·lòfon). A més, per a poder enfinestrar el senyal de veu en trames sonores on la transcripció original no ha marcat periodicitat resulta necessari un nou *ii*) marcat de (*pitch*).

Anàlogament a l'anàlisi exposat en el capítol anterior (apartat 4.2) es presenta un estudi estadístic del corpus que analitza: *i*) la composició del corpus emprat tant a nivell d'al·lòfons com de difonemes, *ii*) la distribució dels diferents paràmetres prosòdics i *iii*) les distribucions dels diferents subcostos emprats, considerant també la seva normalització i

posterior transformació.

5.2.1 Composició fonètica i prosòdica

El corpus *uvig_dav_es* es composa d'un total de 89788 al·lòfons obtinguts a partir de 31 al·lòfons diferents (detallats a la figura 5.2). Aquests 31 al·lòfons es divideixen en 20 al·lòfons sonors i 11 al·lòfons sords. Alhora, els 20 al·lòfons sonors es divideixen en 13 al·lòfons consonàntics, 5 de vocàlics i 2 semivocàlics. Per a més informació, es pot consultar a l'annex D la taula de tipificació fonètica (tipologia, punt d'articulació, mode d'articulació i sonoritat) dels al·lòfons (taula D.1) així com la taula de distribució dels al·lòfons (taula D.2). Segons aquestes dades, els al·lòfons amb més presència dins el corpus són /e/ i /a/ amb una freqüència d'aparicions del 12.62% i 12.38% respectivament, sumant les dues un 25% d'aparicions dels diferents al·lòfons enregistrats. En segon terme apareixen els al·lòfons /o/,/s/,/n/,/r/ i /l/ amb una presència del 9.17%, 7.55%, 5.93%, 5.55% i 5.33% respectivament. En aquesta anàlisi apareix un aspecte destacable: l'al·lòfon vocàlic /u/ té una presència relativament petita en el corpus (1.93%) en contraposició a la resta de vocals, i fins i tot inferior a la semivocal /j/ (2.56%).

A l'hora d'inventariar el corpus en diferents unitats s'introdueixen dues novetats respecte el capítol anterior a efectes d'obtenir un etiquetat més precís:

i) Es diferencien les vocals accentuades (dins d'una síl·laba accentuada) de les vocals àtones (la seva representació va precedida d'un apòstrof (')).

ii) Seguint la notació ToBI (Beckman i Hirschberg, 1994) explicada en l'apartat 2.2.2 s'etiqueten els silencis segons la seva tipologia:

SIL1) Silencis espontanis causats o bé per disfluències (Adell *et al.*, 2006) que introdueixen certa pseudonaturalitat dins la parla o bé s'inclouen per marcar una èmfasi.

SIL2) Silencis dèbils dins la frase (no trenquen el grup d'entonació – p.ex coma, punt i coma).

SIL3) Silencis forts entre frases (delimiten els límits d'un grup entonatiu / *utterance*).

SIL4) Silencis de principi i final de la pista d'àudio enregistrada.

Després d'analitzar la cobertura de les unitats (difonemes obtinguts a través del mòdul de transcripció del grup de recerca (Alías i Iriondo, 2002)), el nombre d'unitats queda fixat en 827 unitats diferents, sumant un total de 88571 unitats enregistrades. A la figura 5.3

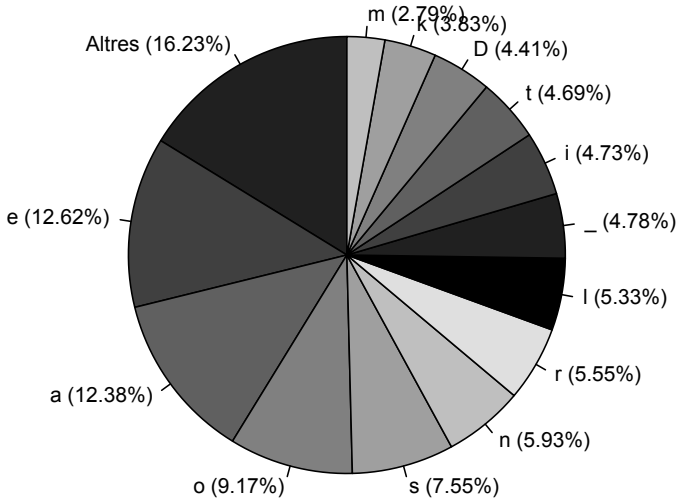


Figura 5.2: Distribució dels diferents al·lòfons en notació SAMPA (Wells *et al.*, 1992) en castellà a través del corpus *uvig_dav_es*.

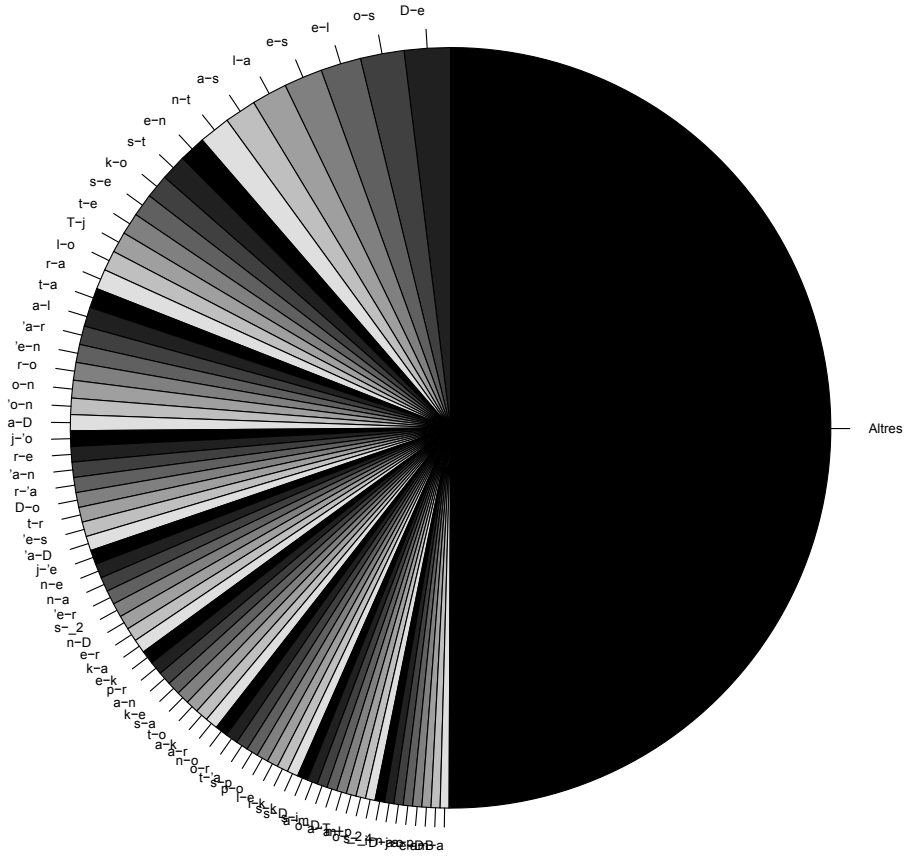


Figura 5.3: Distribució de les diferentes unitats en notació SAMPA (Wells *et al.*, 1992) en castellà a través del corpus *uvig_dav_es*.

s'observa la presència de les unitats més comunes en el corpus. Cal destacar la presència de les 3 vocals principals (/a/,/e/,/o/) en les unitats més representades en el corpus: concretament en 25 de les 28 unitats amb més cobertura (les excepcions són /n-t/,/s-t/ i /T-j/). En l'annex D es pot trobar la taula detallada de cobertura fonètica (taula D.3).

	Max	Min	Mitjana	Mediana	Desviació típica
pitch (Hz)	761	10	100.7285	97	30.742
Δ*pitch* (Hz)	505	-440	0.4946	-1	24.703
energia (RMS)	0.0334	0	0.0045	0.0041	0.0034
Δenergia (RMS)	0.022	-0.0245	0	-2e-04	0.0043
Ritme (*z-scored*)	15.7689	-4.899	0	-0.1294	0.9998
Δ Ritme (*z-scored*)	15.7257	-16.1426	0.002	0.0274	1.4309

Taula 5.1: Estadístiques de primer ordre del corpus *uvig_dav_es*.

A nivell prosòdic, es repeteix l'estudi estadístic sobre la parametrització acústica descrit en l'apartat 4.2.1. No obstant, cal tal tenir en compte que en aquest capítol s'introdueixen els subcostos lingüístics de naturalesa simbòlica (*Independent Feature Formulation* o IFF — veure apartat 2.3.4) a la funció de cost. Aquesta modificació s'explicarà amb més detall a l'apartat 5.3. Per tant, en aquest capítol la descripció estadística de la prosòdia només descriu part dels subcostos amb els que treballa la funció de cost, ja que aquesta es complementa amb subcostos simbòlics.

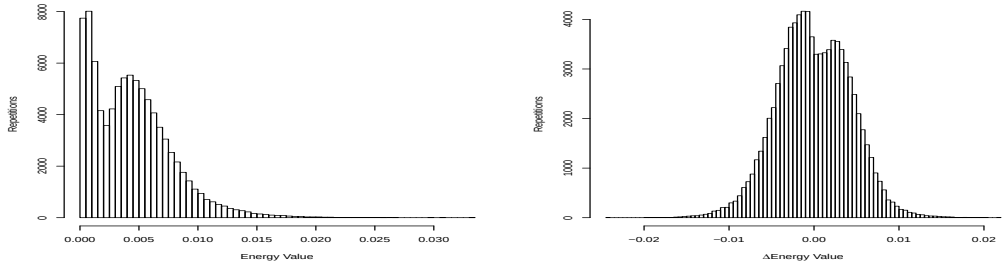
Un altre tret diferencial, és el fet que el *pitch* es marca mitjançant l'eina PRAAT (Boersma i Weenink, 2010) sense supervisió manual i un marge d'anàlisi ampli de $[50 - 600]$ Hz. Concentrant la majoria dels seus valors (97.77%) es concentren en l'interval $[67 - 130]$ Hz.

La taula 5.1 mostra de manera detallada les estadístiques de primer ordre de cada paràmetre prosòdic estudiat sense que a primer cop d'ull se'n pugui treure cap conclusió rellevant respecte la funció de distribució (també dita funció de densitat o pdf) dels diferents valors.

	Asimetria	Curtosi	Test de Lilliefors
pitch	10.5932	145.0143	0.2312
Δ*pitch*	4.8201	115.3584	0.2722
energia	1.0752	1.9433	0.0919
Δenergia	-0.0181	0.2316	0.0215
ritme	1.423	5.8235	0.0881
Δ ritme	-0.0683	3.036	0.0372

Taula 5.2: Estadístiques de segon ordre i proves de normalitat del corpus *uvig_dav_es*.

Llavors, s'analitzen les estadístiques de segon ordre dels paràmetres prosòdics (taula 5.2) per tal d'obtenir més informació sobre la naturalesa de les dades. Observant les dades es pot confirmar l'asimetria de la distribució de les dades sobretot en el *pitch*. També es pot observar com la derivada de l'energia i el ritme (que estan centrades a 0) presenten un nivell d'asimetria baix. Quan s'analitza la curtosi de les dades, s'observa que el *pitch* i la seva derivada presenten una forta concentració al voltant de la mitja. A més, tot i que en menor mesura, el ritme i la seva derivada també presenten una concentració de dades important degut al seu extens fons d'escala. Per tant, aquestes dades indiquen que el corpus té un ritme de producció de la parla constant i no presenta variacions brusques en el contorn entonatiu degut a la seva neutralitat expressiva (Iriondo, 2008) .



(a) Histograma de la distribució de la energia a dins del corpus

(b) Histograma de la distribució de la Δ d'energia a dins del corpus

Figura 5.4: Histogrames de l'energia i la seva derivada a *uvig_esda_es*.

A diferència del corpus en català *url_fer_ct*, ni l'energia ni la seva derivada adopten valors alts de curtosi, però en cap cas prenen valors de curtosi negatius. Aquest fet indica

que les dades no es concentren en cap punt concret de la distribució, però tampoc es tracta d'una distribució multimodal amb més d'una concentració de dades. Aquest fet implica que no resulti evident la distribució bimodal dels al·lòfons dividida en sords i sonors, com passava en *url_fer_ct*. En veure els histogrames de l'energia i la seva derivada (veure figura 5.4) es pot veure com la clara bimodalitat (pdf amb dos màxims) de la distribució de l'energia del corpus *url_fer_ct* queda més diluïda, tot i que segueix adoptant dos màxims ($max_1 \approx 5 \cdot 10^{-4}$ i $max_2 \approx 4 \cdot 10^{-3}$).

L'últim pas de l'anàlisi consisteix en realitzar la prova de normalitat de les dades mitjançant el test de Lilliefors (Lilliefors, 1967) que permet obtenir un índex D que determina la distància d'una distribució respecte la distribució normal (veure apartat A.2 de l'annex). En aquest cas, en disposar de 89788 al·lòfons, el valor mínim d'acceptació per a la prosòdia és $D < 0.00344$ amb una $\alpha = 1\%$. En quan a la seva derivada, es disposa d'una mostra de 88571 valors, per tant el valor crític per a la derivada és $D < 0.003464$. A la taula 5.2 s'observa que cap paràmetre prosòdic, ni tampoc les seves derivades presenten una distribució normal, essent les derivades d'energia i ritme les que més s'hi acosten, mentre que els valors de *pitch* els que menys. En les figures D.1 i D.2 de l'annex D es poden veure amb detall els histogrames i *qqplots* (correlació de les dades respecte la normal) dels diferents paràmetres prosòdics del corpus.

5.2.2 Densitat dels subcostos acústics i millores en la seva normalització

Seguidament, s'estudia la pdf que prenen els diferents subcostos acústics, ja que aquests són els únics que adopten valors continus, deixant de banda els subcostos simbòlics, que adopten valors discrets (veure l'apartat 2.3.2)). Per a realitzar aquest estudi, s'analitzen separadament els subcostos de *target* i els de concatenació, com es pot veure a continuació.

Subcostos de *target*

L'estudi dels subcostos de target segueix les mateixes premisses que l'anàlisi realitzada en l'apartat 4.2.2. En aquest cas, també es considera l'aïllament el mòdul de selecció d'unitats de la resta de mòduls com a pas previ a l'optimització del propi mòdul de selecció d'unitats. Aquest aïllament respecte el mòdul de generació de prosòdia és possible si es disposa d'una prosòdia ideal, exempta d'errors de predicció (veure apartat 4.4.2). Per aconseguir-la s'empren les prosòdies naturals, obtingudes de directament en l'etapa d'etiquetat del corpus com a referència.

Igual que en el corpus *url_fer_ct*, els subcostos acústics de *target* emprats són *i*) el *pitch*

mig (Hz), *ii*) l'energia (RMS), i la*iii*) durada de la unitat (mil·lisegons) considerant sempre la separació del subcost de durada en semifonema esquerre i semifonema dret. El subcost de *pitch* només es té en compte per els semifonemes sonors ja que en els semifonemes sords no té sentit. Cal recordar que la notació seguida és PIT.T pel subcost de *pitch*, ENE.T pel d'energia, DURL.T pel de durada esquerra i DURR.T pel de durada dreta de la unitat. Les estadístiques de primer ordre obtingudes es mostren a la taula 5.3.

Les funcions de densitat (mitjana, desviació típica, etc.) s'obtenen independentment per a cada unitat: per a cada versió U_v d'una unitat U enregistrada en el corpus es bescanvia per la resta de versions U_w (unitats que tenen la mateixa transcripció fonètica) mantenint l'especificació prosòdica original de U_v. Així s'obté una matriu de dimensions $\binom{N}{2} \times M_{targ}$ on N és el nombre de versions que té la unitat i M_{targ} el nombre de subcostos que s'analitzen (en aquest cas, $M_{targ} = 4$).

	Max	Min	Mitjana	Mediana	Desviació típica
PIT.T (Hz.)	562.705	0	16.2924	12.6316	15.7777
ENE.T (RMS)	0.6419	0	0.0741	0.0607	0.0554
DURL.T (ms.)	932.5	0	12.9577	10	16.2391
DURR.T (ms.)	992.5	0	15.557	10	25.6637

Taula 5.3: Estadístiques de primer ordre dels subcostos de *target* pel corpus *uvig_dav_es*.

Els resultats de la taula 5.3 evidencien la diferència de fons d'escala per a cada subcost. A més, el fet que els valors de mitjana i mediana s'ubiquin properes al mínim del seu fons d'escala, essent alhora un valor proper al valor de la desviació típica, denota asimetria en la funció densitat.

En obtenir les estadístiques de segon ordre (taula 5.4) es confirma l'asimetria de les diferents distribucions. A més s'aprecia una alta concentració (curtosi) de dades al voltant de la mitjana especialment en el *pitch*. També s'observa com els subcostos de durada són els que presenten una pdf menys normal ja que presenten valors d'asimetria i curtosi elevats. Els resultats d'aplicar el test de Lilliefors indiquen que cap subcost segueix una distribució normal (el llindar per la grandària de la mostra que tenim és $D < 0.0086$). Els valors obtinguts confirmen els subcostos de durada com els més distants a la distribució normal, seguits dels de *pitch*. A la figura 5.5(a) es poden veure els histogrames detallats per a cada subcost per la unitat /D-e/, que és la unitat més representada en el corpus. Addicionalment a la figura 5.5(b) es pot veure la comparació quartil-quartil respecte una distribució normal.

	Asimetria	Curtosis	Test de Lilliefors
PIT.T	6.5385	128.7027	0.1509
ENE.T	1.3453	2.4451	0.0234
DURL.T	10.1046	229.5974	2.5
DURR.T	7.6754	105.5664	5

Taula 5.4: Estadístiques de segon ordre i proves de normalitat dels subcostos de *target* del corpus *uvig_dav_es*.

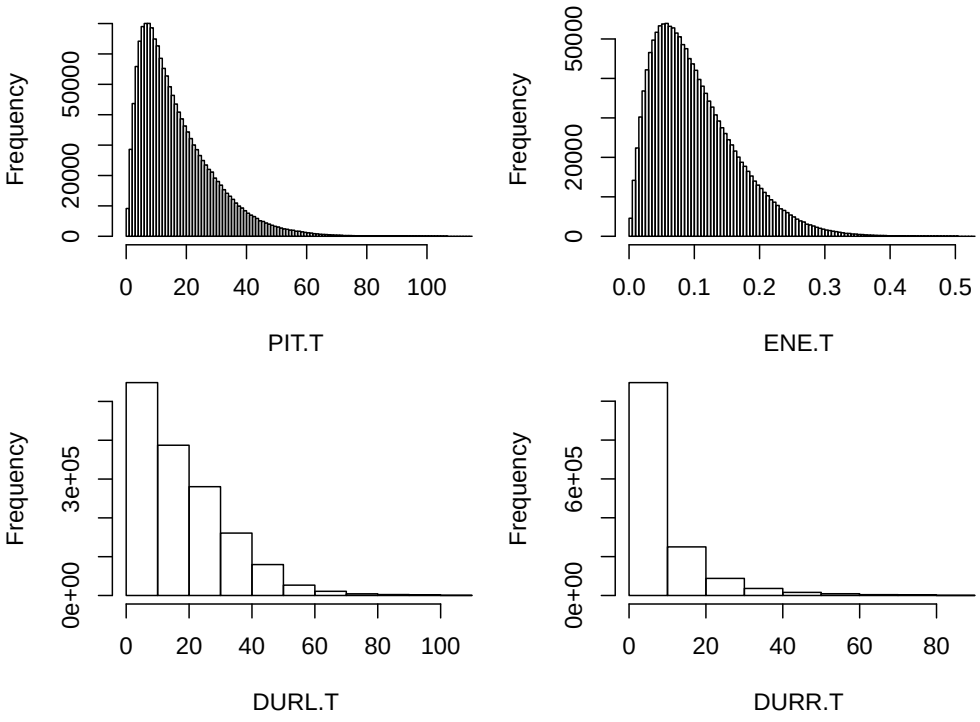
Subcostos de concatenació

A diferència dels subcostos de *target*, els subcostos de concatenació s'obtenen a través de la diferència de paràmetres acústics en el punt de concatenació entre dues unitats (veure l'apartat 2.3.3).

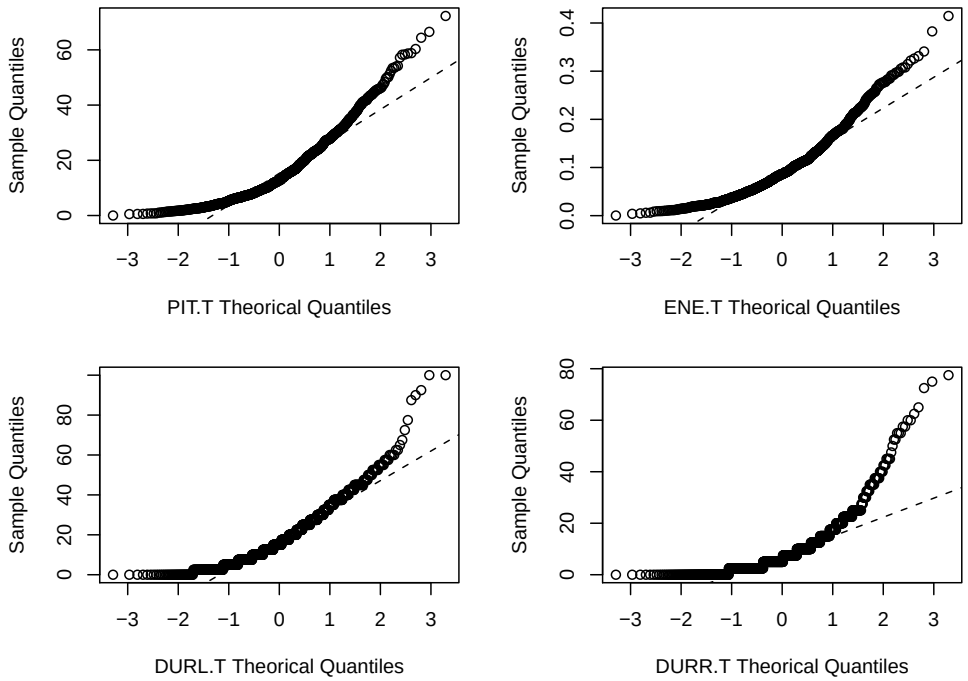
Els subcostos de concatenació escollits són: *i*) la discontinuïtat de *pitch* (Hz), *ii*) d'energia (RMS), i *iii*) la discontinuïtat espectral calculada a partir dels paràmetres MFCC. De la mateixa manera que en el corpus *url_fer_ct*, el subcost MFCC es calcula com la distància dels 12 primers coeficients mel-cepstrals en el punt de concatenació (exceptuant el C_0, que és l'energia) juntament amb les seves derivades. En les concatenacions on intervenen unitats sordes no es calcula el subcost de concatenació de *pitch*, donat que les unitats sordes no tenen assignat valor de F_0. La notació seguida pels subcostos de concatenació és: PIT.C per la discontinuïtat de *pitch*, ENE.C per la d'energia i MFC.C per l'espectral.

Pels subcostos de concatenació, les estadístiques també s'obtenen de manera independent per a cada unitat (unitat prèvia a la concatenació). Per cadascuna de les variants U_i d'una unitat U del corpus es calculen els subcostos derivats de concatenar-la amb totes les possibles variants W_j, on W representa totes les unitats que poden ser concatenades amb U_i (el semifonema dret d'U_i és el mateix que el semifonema esquerre de W). Com a resultat s'obté una matriu $N' \times M_{conc}$ on N' és el nombre de concatenacions possibles i M_{conc} és el nombre de paràmetres analitzats ($M_{conc} = 3$).

Les estadístiques de primer ordre dels subcostos de concatenació es mostren a la taula 5.5. En aquest cas, els valors estadístics obtinguts són estimats ja que degut a la seva mida ($n \approx 396 \cdot 10^6$ valors per cada subcost), s'han hagut de submostrejar uniformament els subcostos per poder-ne obtenir els resums estadístics de manera computacional.



(a) Histogrames.



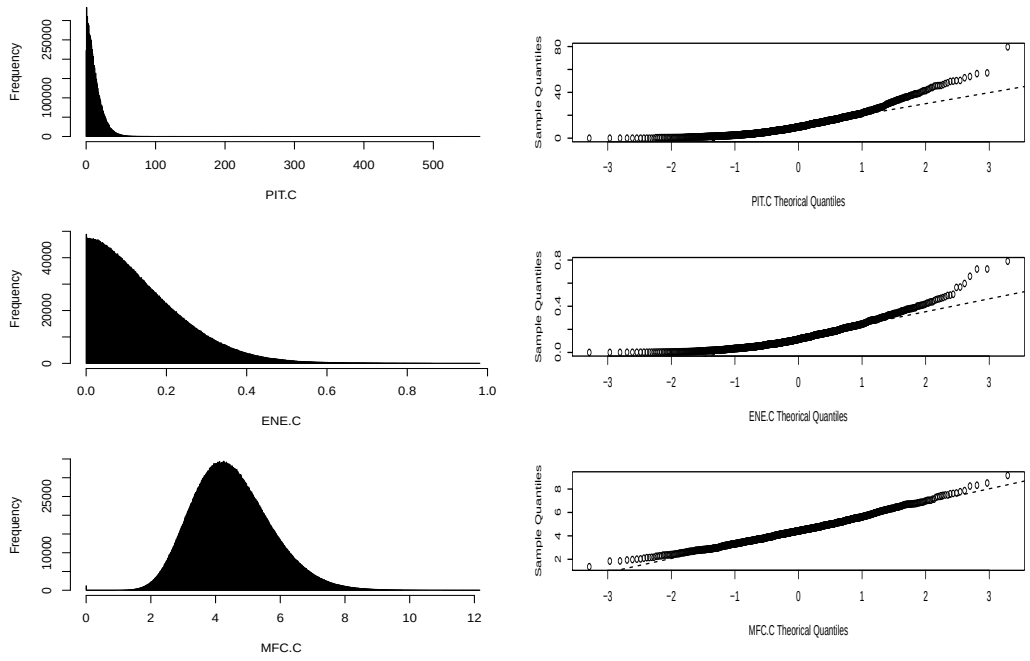
(b) Comparació quartil-quartil.

Figura 5.5: Histogrames i comparació quartil-quartil (*qqplot*) respecte la distribució normal dels subcostos de *target* per a la unitat /D-e/.

	Max	Min	Mitjana	Mediana	Desviació típica
PIT.C (Hz)	572.203	0	14.0529	11.0745	13.7412
ENE.C (RMS)	1.1228	0	0.0914	0.0593	0.0948
MFC.C (MFCC val)	16.1506	0	5.0401	4.9040	1.4955

Taula 5.5: Estadístiques de primer ordre dels subcostos de concatenació corpus *uvig_dav_es*.

Una altra vegada, s'observa la necessitat de normalitzar els subcostos a un fons d'escala [0,1] per tal que els subcostos puguin ser comparats entre sí. Es torna a observar una asimetria en la distribució de dades pel fet d'obtenir diferents valors de mediana i mitjana. Tanmateix aquests valors són molt més pròxims al mínim que al màxim del seu fons d'escala. En quant a les desviacions típiques, aquestes adopten gairebé la mateixa magnitud que la mediana (a excepció de MFC.C).



(a) Histogrames dels diferents subcostos de concatenació per a la unitat /D-e/.

(b) Comparació quartil-quartil dels subcostos de concatenació respecte la distribució normal per a la unitat /D-e/.

Figura 5.6: Histogrames i comparació quartil-quartil (*qqplot*) dels subcostos de concatenació per a la unitat /D-e/.

Analitzant les estadístiques de segon ordre de la taula 5.6 (també estimades degut al submostrejat), s'observen dos grups clarament diferenciats en quant a asimetria i curtosi. Per una banda s'observa que el *pitch* i l'energia presenten distribucions similars i mentre que, el subcost de MFC.C presenta una distribució més normal. Altra vegada cap subcost passa el test de Lilliefors ($D < 0.00016$) per cap unitat tot i que segons els valors aproximats obtinguts s'aprecia com el subcost de MFC.C és el que s'acosta més a una distribució normal i el subcost de PIT.C el que menys normalitat presenta. En quant a l'asimetria i la concentració de les dades, s'observa que el subcost de *pitch* presenta una concentració de dades molt alta (anàlogament al que s'observava pel subcost de *pitch* de *target*). A més, es pot observar com el subcost de MFC.C presenta uns valors d'asimetria i curtosi més baixos. A la figura 5.6(a) es poden veure els histogrames dels diferents subcostos de concatenació. Addicionalment a la figura 5.6(b) es pot veure la comparació quartil-quartil dels subcostos de concatenació respecte a la distribució normal.

	Asimetria	Curtosis	Test de Lilliefors
PIT.C	7.3142	173.3681	0.1532
ENE.C	1.8301	4.6189	0.1676
MFC.C	0.4835	0.1469	0.0365

Taula 5.6: Estadístiques de segon ordre i proves de normalitat dels subcostos de concatenació del corpus *uvig_dav_es*.

Normalització dels subcostos

En l'apartat 4.2.2 s'ha explicat la necessitat d'escalar les dades a un fons d'escala comú, així com el procés seguit per a transformar-les per assolir normalitat i encabir-les en un model de regressió lineal (veure apartat 2.4.2). En resum, per a poder extreure models de regressió a partir de les dades, les seves funcions de densitat de probabilitat s'han de transformar segons les condicions següents: *i*) assegurar la seva linealitat (creixement dels subcostos proporcionalment lineal respecte la degradació espectral – veure apartat 3.2.4), *ii*) assolir normalitat i *iii*) estabilitzar la variància de les mateixes (Chatterjee i Hadi, 2006).

Per assolir els objectius *ii* i *iii*, fins al moment, s'han estudiat les funcions de normalització / transformació següents:

$$SC_i^{\text{Z-SCORED}} = \frac{SC_i - \overline{SC}}{\sigma_{SC}} \tag{5.1}$$

$$SC_i^{\text{MAX-MIN}} = \frac{SC_i - \min(SC)}{\max(SC) - \min(SC)} \tag{5.2}$$

$$SC_i^{\text{SIGMOID}^2} = 1 - e^{-\left(\frac{SC_i}{\sigma_{SC}}\right)^2} \tag{5.3}$$

$$SC_i^{\text{SIGMOID}} = 1 - e^{-\left(\frac{SC_i}{\sigma_{SC}}\right)} \tag{5.4}$$

on SC_i^k representa la normalització del valor i de la distribució segons la transformació $k = \{\text{Z} - \text{SCORED}, \text{MAX} - \text{MIN}, \text{SIGMOID}^2, \text{SIGMOID}\}$. $\overline{SC}$, $\max(SC)$, $\min(SC)$ i σ_{SC} representen la mitjana, el valor màxim, el valor mínim i la desviació típica de la distribució respectivament.

Malgrat que la transformació SC^{SIGMOID} proporciona una millora notable en la normalitat de la pdf (veure apartat 4.2.2), la saturació en la part alta de l'escala impedeix assolir una normalitat òptima. Tukey (1957) va proposar l'anomenada *escala de transformacions* que suggereix el tipus de distribució que s'ha d'aplicar, segons sigui l'intensitat de l'asimetria de la distribució o la localització dels *outliers*. Aquestes transformacions es coneixen com a transformadors de potència, i es separen entre les que corregeixen l'asimetria positiva i les que corregeixen l'asimetria negativa. Com que els subcostos de la funció de cost analitzats presenten asimetria positiva, és necessari d'aplicar transformacions de a família $\sqrt{X}$,$log(X)$,$\frac{1}{X}$,... (Tukey, 1957). Per aquest motiu en aquest capítol s'implementen dues funcions del conjunt de transformació de Tukey (1957) amb l'objectiu de millorar la funció de densitat (pdf):

$$SC_i^{\text{LOG}} = \frac{\log(1 + SC_i) - \log(1 + \min(SC))}{\log(1 + \max(SC)) - \log(1 + \min(SC))} \tag{5.5}$$

$$SC_i^{\text{SQRT}} = \frac{\sqrt{SC_i} - \min(\sqrt{SC})}{\max(\sqrt{SC}) - \min(\sqrt{SC})} \tag{5.6}$$

Aquestes transformacions pertanyen a la família de funcions que corregeixen l'asimetria positiva ($\sqrt{X}$,$log(X)$,$\frac{1}{X}$...) d'una pdf. La transformació SC^{LOG} és una transformació logarítmica i la transformació SC^{SQRT} és una transformació d'arrel. La figura 5.7 mostra el comportament de les noves funcions, juntament amb les funcions sigmoide, en el domini de valors que pren el subcost PIT.T per a la unitat /D-e/.

	Desviació típica	Asimetria (*skewness*)	Curtosi	Test de Lilliefors
PIT.T	0.2691 (*0.123*)	-0.2694 (*6.5385*)	-1.0239 (*128.7027*)	0.0655 (*0.1509*)
ENE.T	0.2280 (*0.1294*)	-0.6558 (*1.3453*)	-0.447 (*2.4451*)	0.0837 (*0.0974*)
DURL.T	0.2882 (*0.1293*)	-0.374 (*10.1046*)	-0.9361 (*229.5974*)	0.0864 (*0.2125*)
DURR.T	0.2939 (*0.1409*)	-0.1817 (*7.6754*)	-1.0748 (*105.5664*)	0.0692 (*0.2722*)
PIT.C	0.2749 (*0.0793*)	-0.235 (*7.3142*)	-1.0745 (*173.3681*)	0.0648(*0.1532*)
ENE.C	0.2812 (*0.106*)	-0.3045 (*1.8301*)	-1.0709 (*4.6189*)	0.0731(*0.1676*)
MFC.C	0.0430 (*0.1091*)	-3.9982 (*0.4835*)	51.6388 (*0.1469*)	0.1719 (*0.0365*)

Taula 5.7: Estadístiques de segon ordre obtingudes després d'aplicar la transformació exponencial (sigmoide lineal) en els subcostos del corpus *uvig_dav_es*. Entre parèntesi hi ha els valors obtinguts amb la transformació *max-min*.

	Desviació típica	Asimetria (*skewness*)	Curtosi	Test de Lilliefors
PIT.T	0.4048 (*0.123*)	0.5319 (*6.5385*)	-0.0038 (*128.7027*)	0.0479 (*0.1509*)
ENE.T	0.3807 (*0.1294*)	0.2046 (*1.3453*)	-0.5257 (*2.4451*)	0.0281 (*0.0974*)
DURL.T	0.4214 (*0.1293*)	0.3756 (*10.1046*)	-0.3405 (*229.5974*)	0.0559 (*0.2125*)
DURR.T	0.4268 (*0.1409*)	0.5456 (*7.6754*)	-0.259 (*105.5664*)	0.0761 (*0.2722*)
PIT.C	0.3982 (*0.0793*)	0.5467 (*7.3142*)	0.2783 (*173.3681*)	0.0527 (*0.1532*)
ENE.C	0.4178 (*0.106*)	0.418 (*1.8301*)	-0.4262 (*4.6189*)	0.0513 (*0.1676*)
MFC.C	0.2234 (*0.1091*)	-0.2741 (*0.4835*)	0.2267 (*0.1469*)	0.0187 (*0.0365*)

Taula 5.8: Estadístiques de segon ordre obtingudes després d'aplicar la transformació logarítmica en els subcostos del corpus *uvig_dav_es*. Entre parèntesi hi ha els valors obtinguts amb la transformació *max-min*.

	Desviació típica	Asimetria (*skewness*)	Curtosi	Test de Lilliefors
PIT.T	0.4419 (*0.123*)	0.5355 (*6.5385*)	0.8003 (*128.7027*)	0.0317 (*0.1509*)
ENE.T	0.4053 (*0.1294*)	0.268 (*1.3453*)	-0.2256 (*2.4451*)	0.024 (*0.0974*)
DURL.T	0.4857 (*0.1293*)	0.0825 (*10.1046*)	0.1617 (*229.5974*)	0.0558 (*0.2125*)
DURR.T	0.4935 (*0.1409*)	0.234 (*7.6754*)	0.0427 (*105.5664*)	0.0547 (*0.2722*)
PIT.C	0.4403 (*0.0793*)	0.5772 (*7.3142*)	2.1016 (*173.3681*)	0.0253 (*0.1532*)
ENE.C	0.4565 (*0.106*)	0.3567 (*1.8301*)	-0.0946 (*4.6189*)	0.0252 (*0.1676*)
MFC.C	0.2694 (*0.1091*)	-0.0719 (*0.4835*)	0.2313 (*0.1469*)	0.005 (*0.0365*)

Taula 5.9: Estadístiques de segon ordre obtingudes després d'aplicar la transformació d'arrel en els subcostos del corpus *uvig_dav_es*. Entre parèntesi hi ha els valors obtinguts amb la transformació *max-min*.

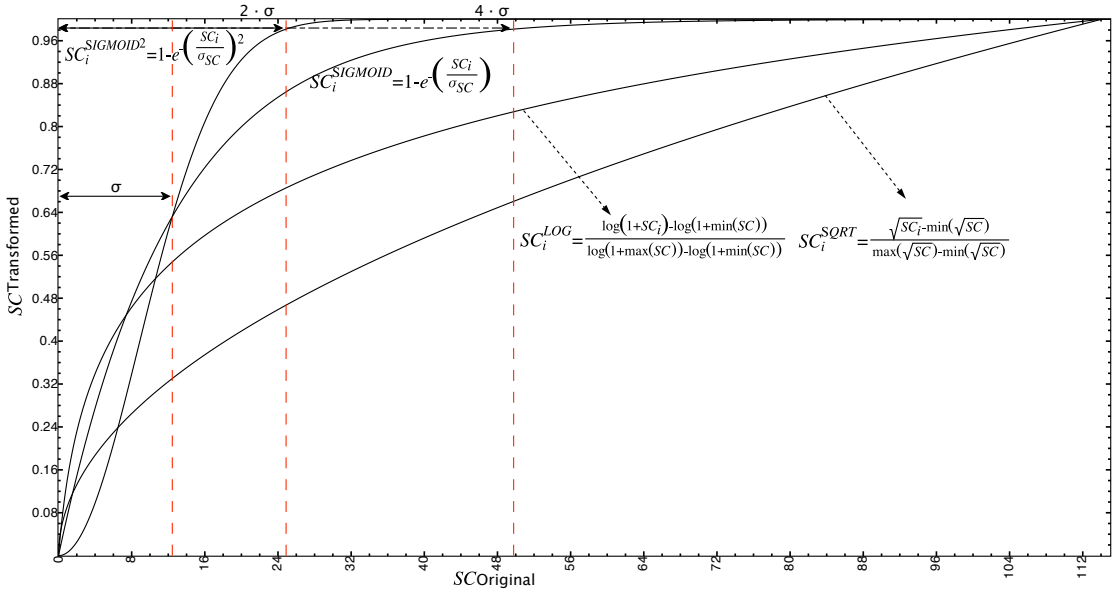


Figura 5.7: Funcions de transformació aplicades sobre el subcost PIT.T de la unitat /D-e/.

Per tal d'analitzar els efectes de la funció sigmoide sobre les dades originals, es tornen a obtenir les estadístiques de segon ordre un cop aplicades les diferents transformacions estudiades. Els resultats es detallen en les taules 5.7, 5.8 i 5.9.

La transformació SC^{SIGMOID}, malgrat millorar la normalitat de les dades, en aquest corpus empitjora l'estabilització de la desviació típica pels diferents subcostos. Així, es passa d'una variabilitat del 17.55% obtinguda amb $SC^{\text{MAX}-\text{MIN}}$ a un 37.28%. Aplicant les noves funcions de Tukey (1957), aquesta variabilitat s'acosta més als valors originals ja que s'obtenen valors de 18.76% amb SC^{LOG} i 17.72% amb SC^{SQRT}. Respecte l'asimetria i la curtosi no s'aprecien diferències destacables més enllà d'homogeneitzar el subcost de MFC.C amb la resta de subcostos. En quant a la normalitat de les dades segons el test de Lilliefors, s'aprecia una clara millora en tots els subcostos al obtenir un valor D més petit. No obstant això, cap dels estadístics D aconsegueix superar el llindar de normalitat de les dades ($D = 1.97 \cdot 10^{-3}$)

Aquest estudi sobre la normalització dels subcostos es completa analitzant amb detall la normalitat dels subcostos per a cada unitat considerant com a mostra les 100 unitats més poblades del corpus. Per tant, per cadascuna d'aquestes unitats s'avalua el seu índex de normalitat D de Lilliefors segons la normalització sense transformació (*max-min*), i les diferents funcions de transformació emprades. El resultat d'aquest estudi es mostra a la

figura 5.8 i a la taula 5.10.

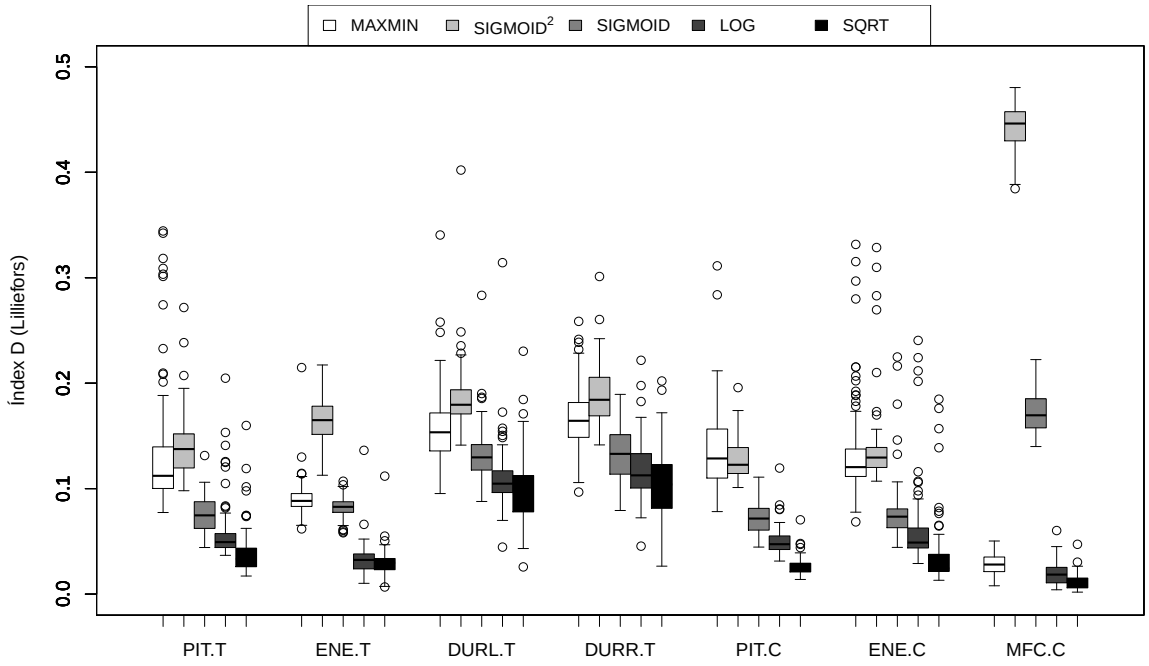


Figura 5.8: Resultat del test de Lilliefors aplicat als subcostos prosòdics de les 100 unitats més representades en el corpus *uvig_dav_es* segons les diferents normalitzacions estudiades.

Segons els resultats obtinguts, es confirmen els aconseguits en l'estudi global dels subcostos a nivell d'unitat: l'índex de normalitat D millora en transformar els subcostos segons les funcions de Tukey (1957). Entre les funcions proposades per, la millor normalitat s'assoleix amb la funció SQRT, la qual obté el millor índex D per a tots els subcostos acústics analitzats.

Per últim, s'analitza si les diferències són significatives mitjançant el test de t de Student per parelles segons la correcció de Bonferroni (necessària quan es comparen més de dos grups de dades) (Hochberg, 1988). En aquest estudi, la majoria de diferències entre els mètodes de transformació són clarament significatives ($p < 2 \cdot 10^{-16}$ excepte en les comparatives que s'indiquen a continuació: *i*) $\text{PIT.T}^{LOG} = \text{PIT.T}^{SIGMOID}$ ($p = 0.174$), *ii*) $\text{PIT.T}^{LOG} = \text{PIT.T}^{SQRT}$ ($p = 0.098$), *iii*) $\text{DURR.T}^{LOG} = \text{DURR.T}^{SIGMOID}$ ($p = 0.072$), *iv*) $\text{DURR.T}^{LOG} = \text{DURR.T}^{SQRT}$ ($p = 0.05$) i *v*) $\text{ENE.C}^{LOG} = \text{ENE.C}^{EXP}$ ($p = 0.06566$)

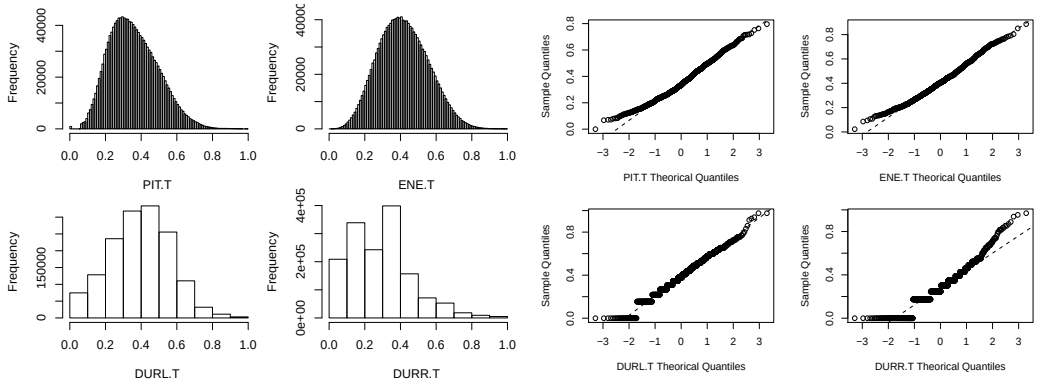
Per tant, la funció de transformació SQRT és la que s'utilitzarà com a funció de nor-

Mitjana	MAX-MIN	SIGMOID^2	SIGMOID	LOG	SQRT
PIT.T	0.1341	0.1395	0.0742	0.0573	0.0392
ENE.T	0.0909	0.1651	0.082	0.0325	0.0297
DUR.L	0.1579	0.1849	0.1309	0.1093	0.0983
DUR.R	0.1678	0.1892	0.1303	0.1183	0.1058
PIT.C	0.1391	0.1277	0.0715	0.0505	0.0265
ENE.C	0.1346	0.1373	0.0768	0.0604	0.0356
MFC.C	0.0277	0.4423	0.1722	0.0193	0.0115

Taula 5.10: Resultat d'aplicar el test de Lilliefors en els subcostos de les 100 unitats més representatives de corpus *uvig_dav_es* segons les diferents normalitzacions estudiades.

malització en les contribucions d'aquest capítol aplicades sobre les unitats del corpus *uvig_dav_es*.

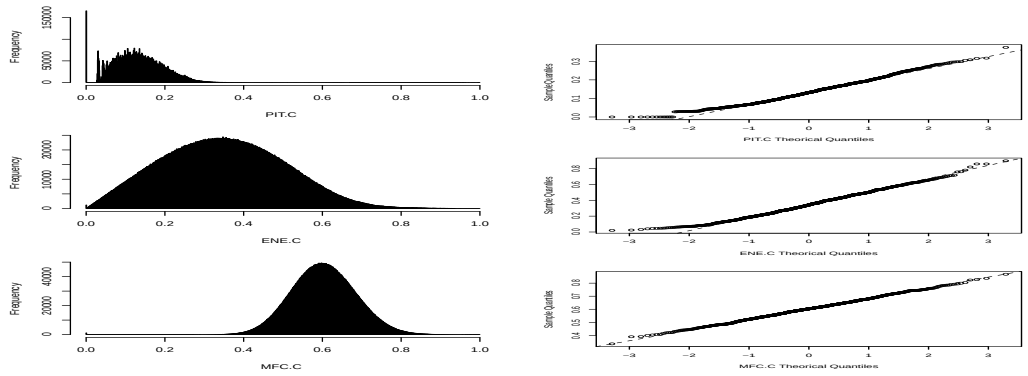
Les funcions de densitat finals (SQRT) per la unitat /D-e/ es mostren a les figures 5.9 i 5.10, on es presenten detallades les funcions de densitat dels valors dels subcostos de *target* (figura 5.9(a)), la comparació quartil-quartil (*qqplot*) dels mateixos respecte la distribució normal (figura 5.9(b)), la densitat de valors dels subcostos de concatenació (figura 5.10(a)) i la respectiva comparació quartil-quartil (*qqplot*) respecte la distribució normal (figura 5.10(b)).



(a) Histogrames dels diferents subcostos de *target*.

(b) Comparació quartil-quartil dels subcostos de *target* transformats segons la normalització d'arrel respecte la distribució normal.

Figura 5.9: Histogrames i comparació quartil-quartil (*qqplot*) dels subcostos de *target* transformats per a la unitat /D-e/ *uvig_dav_es*.



(a) Histogrames dels diferents subcostos de concatenació.

(b) Comparació quartil-quartil dels subcostos de concatenació transformats segons la normalització d'arrel respecte la distribució normal.

Figura 5.10: Histogrames i comparació quartil-quartil (*qqplot*) dels subcostos de concatenació transformats per a la unitat */D-e/* del corpus *uvig_dav_es*.

5.3 Introducció dels subcostos lingüístics

5.3.1 Motivació

En l'apartat 2.3.4 s'ha exposat l'aparent contradicció existent en emprar subcostos acústics (ASF) o lingüístics (IFF) en la funció de *target*. No obstant, quan s'empren subcostos lingüístics la cerca no conté els errors previs d'altres mòduls (etiquetat del corpus – predicció de prosòdia) però sí poca precisió a l'hora de discriminar les unitats candidates (ambigüitat). En canvi, aquesta ambigüitat queda resolta en l'especificació acústica (més precisa – valors continus) però que comporta introduir errors en l'especificació de la seqüència d'unitats desitjada.

La recerca realitzada els últims anys prioritza els subcostos lingüístics sobre els subcostos acústics sota la consideració que l'espai acústic (ASF) simplement és una transformació de l'espai lingüístic (IFF) però que incorpora errors (Clark *et al.*, 2005; Colotte i Beaufort, 2005; Tihelka, 2005), alhora que resulta més complexe en haver d'implementar un mòdul de predicció de prosòdia (Taylor, 2009). No obstant, els mateixos autors reconeixen que reduir l'ambigüitat en la cerca augmentant les característiques lingüístiques provoca un augment la dispersió de les dades, evidenciant la importància d'un estudi rigorós sobre ponderació òptima dels subcostos (Strom i King, 2008). És important destacar que en el

treball de (Strom i King, 2008) s'intenta definir una nova funció de cost més avançada que la de (Hunt i Black, 1996) realitzant dos mapatges a tres bandes (lingüístic → acústic, lingüístic → perceptiu). Malgrat que obtenen conclusions interessants (l'accent terciari i el context no provoquen grans diferències perceptives), cal notar que els seus resultats no són generalitzables ja que només realitzen l'estudi per dos difonemes concrets amb una sola frase portadora.

Resulta evident que s'ha de realitzar un judici previ en el disseny de la funció de cost per tal d'assolir una selecció d'unitats òptima. No obstant, recentment predominen les aproximacions híbrides que consideren treballar juntament amb la informació acústica i la informació lingüística, (Toda *et al.*, 2006; Campillo i Rodríguez-Banga, 2006; Bonafonte *et al.*, 2008). En aquest sentit es pot pensar que emprant només uns o altres de manera exclusiva, és perd part de la informació important per realitzar una bona selecció d'unitats.

En aquest sentit, tal com s'ha esmentat a l'apartat 2.3.4, la discussió sobre les característiques a considerar en la funció de cost (sobretot de *target*) no està tancada i necessita d'un estudi més exhaustiu. Cal tenir en compte que el paradigma de (Hunt i Black, 1996) no limita la tipologia o naturalesa de les característiques a emprar ni tampoc la seva representació (contínua, discreta...).

És per aquest motiu que en aquest capítol es faran competir els subcostos acústics (ASF) emprats fins ara (Alías (2006) i capítol 4) amb els subcostos lingüístics (IFF) per tal que sigui l'usuari qui determini la importància de cada característica en la selecció d'unitats per cada tipus d'unitat considerada, permetent en tot moment que sigui l'usuari que determini la importància de les diferents tipologies dins d'un model híbrid.

A més, una de les avantatges de la metodologia d'ajust de pesos basada en aiGA que es proposa en aquesta tesi doctoral és que permet determinar la importància dels pesos de manera interactiva per part de persones no expertes. A més, segons l'agrupament mitjançant CART exposat en l'apartat 4.4.4 s'aconsegueix respectar la importància dels subcostos segons l'especificitat de cada unitat.

5.3.2 Subcostos incorporats

A l'hora de definir quins subcostos s'incorporen en la funció de cost es parteix del treball de Clark *et al.* (2007) emprats en la plataforma de síntesi *Multisyn*. La elecció d'aquest treball ve donat perque *Multisyn*, al ser de codi obert, és una plataforma àmpliament acceptada en l'entorn dels CTP-SU i ofereix un marc de comparació base per altres sistemes CTP-SU (Taylor, 2009).

Aquests subcostos es detallen a continuació:

1. PosInEG.L: Adequació de la posició dins el grup d'entonació (0: Mateixa posició / 0.5: Posició veïna / 1: Posició oposada on les posicions són inicial, mig i final)

2. PosInWord.L: Adequació de la posició dins la paraula (0: Mateixa posició / 0.5: Posició veïna / 1: Posició oposada on les posicions son inicial, mig i final)

3. PosInSyl.L: Adequació de la posició dins la síl·laba (0: Mateixa posició / 0.5: Posició veïna / 1: Posició oposada on les posicions son inicial, mig i final)

4. Prev.L: Similaritat del fonema anterior a la unitat (0: Mateix fonema / 1: Diferent fonema)

5. Next.L: Similaritat del fonema posterior a la unitat. (0: Mateix fonema / 1: Diferent fonema)

6. POS.L: *Part-of-Speech* o categoria gramatical (0: Mateixa categoria / 1: Diferent categoria)

7. Stress.L: Fonema en síl·laba accentuada (0: Coincideix accentuació síl·labes / 1: No coincideix)

Tal com s'ha dit (apartat 2.3.4), el treball de Clark *et al.* (2007) sosté que només emprant característiques lingüístiques (amb els seus valors discrets o simbòlics) resulten suficients per obtenir una bona selecció d'unitats. De totes maneres, admeten que en situacions on es requereixi un estil diferent que l'enunciatiu (també anomenat declaratiu o neutre) s'hauria de desenvolupar un mòdul que tractés almenys l'entonació i l'accentuació.

En el treball de Clark *et al.* (2007) es pondera la importància de cada subcost lingüístic a mà, basant-se en la intuïció del dissenyador. Parteixen senzillament de les premisses següents: l'accentuació (*stress*) i la posició en el grup d'entonació són generalment considerades com les característiques més importants en la selecció d'unitats. En aquest sentit només empren la prosòdia a efectes d'eliminar unitats del corpus que s'hagin etiquetat com *outliers* en termes de duració o *pitch*. La ponderació emprada es mostra a la taula 5.11.

Subcost	*Pes*
PosInEG.L	15
Stress.L	10
POS.L	6
PosInSyl.L	5
PosInWord.L	5
Prev.L	4
Next.L	3

Taula 5.11: Ponderació dels subcostos definida a Clark *et al.* (2007). Els valors finals de pesos es normalitzen posteriorment tal que $\sum_i w_i = 1$.

5.4 Nova precisió en l'ajust dels pesos automàtic: ajust a partir de subunitats contextualitzades

5.4.1 Motivacions

El fet d'introduir informació lingüística en l'estructura formal que defineix la unitat permet reformular el nivell òptim d'ajust en la selecció d'unitats (veure apartat 4.4). Fins al moment, es considerava el nivell d'unitat com el nivell de detall màxim en l'ajust de pesos (Hunt i Black, 1996; Campbell i Black, 1997). Aquest fet infereix implícitament que només existeix una sola combinació de pesos òptima per als diferents contexts en els que pot trobar la unitat, descartant conceptes específics tals com les unitats veïnes, l'accentuació o la ubicació. A Strom i King (2008) s'exposa que no es pot inferir una certa combinació de pesos com a bona per a totes les realitzacions d'una unitat degut a que la seva percepció varia en funció del seu context. De fet en el treball presentat per (Black i Taylor, 1997*a*) es realitza una primera aproximació a l'agrupament de les unitats segons un criteri acústic (distància cepstral) que alhora considera el seu context fonètic, lingüístic o accentual.

Conseqüentment, es pot tipificar un nou esquema d'ajust dels pesos (figura 5.11). Val a dir que aquest nou nivell de detall no impedeix ajustar els pesos a nivell de clúster, tal i com ja s'ha exposat en el capítol anterior (apartat 4.4); només cal modificar l'element bàsic d'ajust passant d'unitat a subunitat contextualitzada. Per tant, l'ajust basat en l'agrupament de patrons de pesos dins el corpus es pot adaptar considerant el context de la unitat juntament amb la seva especificitat fonètica.

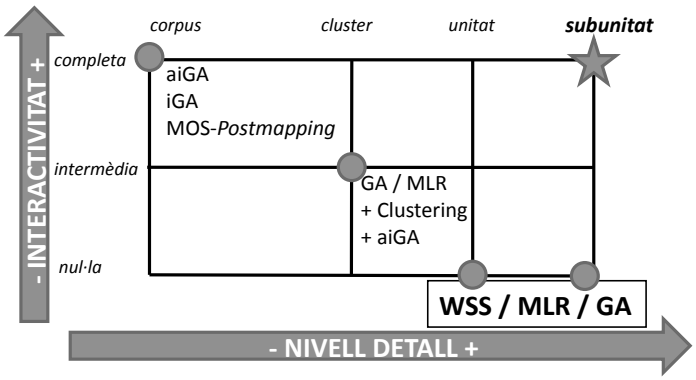


Figura 5.11: Esquema que representa els diferents nivells de detall en l'ajust de pesos en funció de l'interactivitat que ofereix el mètode d'ajust (l'estrella mostra l'ajust desitjat).

5.4.2 Nou escenari: nivell de subunitat

Abans d'explicar els canvis que suposa aquest nou nivell d'ajust de pesos en el procés d'optimització convé repassar l'esquema actual d'ajust de pesos des d'un punt de vista del nivell d'ajust: quan s'ajusten els pesos a nivell d'unitat es considera una sola matriu que engloba totes les especificacions possibles d'entrada (figura 5.12(a)). A partir d'aquesta matriu es calcula una sola combinació de pesos per tota la unitat sense tenir en compte si aquests prenen diferents valors en funció del context lingüístic o fonètic en el que es troba (figura5.12(c)).

El principal canvi que comporta treballar a nivell de subunitat és passar d'una sola combinació de pesos (vector) per unitat a M combinacions (vectors) per unitat, on M és el nombre de versions enregistrades de la mateixa unitat en el corpus. Tal com es pot veure en la figura 5.12(b) la matriu original queda subdivida en M submatrius que relacionen les distàncies automàtiques (cepstrals a (Hunt i Black, 1996)) amb els subcostos segons l'especificació de *target* donada per la versió pertinent de la unitat. Llavors, els pesos calculats s'ubiquen en la taula global (figura 5.12(d)) que és la que servirà per detectar patrons de pesos tal i com s'explicarà en l'apartat 5.5.

Target	Candidata	Dist. Cepstr.	$SC^{PIT.T}$	$SC^{ENE.T}$	$SC^{DURL.T}$	…	$SC^{MFC.C}$
$D-e_1$	$D-e_{1,1}$	18.98	0.28	0.29	0.49	…	0.67
$D-e_1$	$D-e_{1,2}$	21.03	0.35	0.41	0.41	…	0.44
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
$D-e_2$	$D-e_{2,1}$	18.98	0.28	0.29	0.49	…	0.67
$D-e_2$	$D-e_{2,2}$	21.03	0.35	0.41	0.41	…	0.44
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
$D-e_M$	$D-e_{M,1}$	25.51	0.62	0.53	0.41	…	0.47
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
$D-e_M$	$D-e_{M,20}$	30.10	0.61	0.50	0.22	…	0.41

(a) Taula de subcostos per la unitat /D-e/

Target	Candidata	Dist. Cepstr.	$SC^{PIT.T}$	$SC^{ENE.T}$	$SC^{DURL.T}$	…	$SC^{POS.L}$
$D-e_1$	$D-e_{1,1}$	18.98	0.28	0.29	0.49	…	0
$D-e_1$	$D-e_{1,2}$	21.03	0.35	0.41	0.41	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
$D-e_1$	$D-e_{1,20}$	25.49	0.47	0.52	0.44	…	1
$D-e_2$	$D-e_{2,1}$	23.53	0.39	0.50	0.38	…	0
$D-e_2$	$D-e_{2,2}$	24.50	0.24	0.45	0.46	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
$D-e_2$	$D-e_{2,20}$	38.13	0.19	0.41	0.31	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
$D-e_M$	$D-e_{M,1}$	25.51	0.62	0.53	0.41	…	1
$D-e_M$	$D-e_{M,2}$	26.08	0.67	0.60	0.41	…	1
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
$D-e_M$	$D-e_{M,20}$	30.10	0.61	0.50	0.22	…	1

(b) Taula de subcostos per totes les subunitats de la unitat /D-e/

Target	$W^{PIT.T}$	$W^{ENE.T}$	$W^{DURL.T}$	…	$W^{MFC.C}$
$D-e$	0.45	0.03	0.10	…	0.13
$o-s$	0.44	0.02	0.02	…	0.01
⋮	⋮	⋮	⋮	⋱	⋮
$a-'u$	0.51	0.06	0.18	…	0

(c) Taula de pesos automàtics per tot el corpus a nivell d'unitat

Target	PosInEG.L	PREV	…	isSylStressed	$W^{PIT.T}$	$W^{ENE.T}$	…	$W^{POS.L}$
$D-e_1$	CENTRE	e	…	FALSE	0.51	0.32	…	0.00
$D-e_2$	END	a	…	TRUE	0.40	0.12	…	0.02
⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
$D-e_{M_{D-e}}$	INIT	r	…	FALSE	0.11	0.23	…	0.14
$o-s_1$	END	l	…	FALSE	0.52	0.00	…	0.12
$o-s_2$	CENTRE	t	…	TRUE	0.53	0.10	…	0.00
⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
$o-s_{M_{o-s}}$	CENTRE	r	…	FALSE	0.71	0.00	…	0.00
⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
$a-'u_1$	CENTRE	SIL4	…	HALF	0.00	0.30	…	0.04
$a-'u_2$	CENTRE	e	…	HALF	0.25	0.12	…	0.000
⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
$a-'u_{M_{a-'u}}$	CENTRE	s	…	HALF	0.00	0.03	…	0.07

(d) Taula de pesos automàtics per tot el corpus a nivell de subunitat

Figura 5.12: Taules de subcostos i pesos obtingudes a nivell d'unitat (esquerra) i subunitat (dreta).

Nom	Descripció
LEFT	Al·lòfon de l'esquerra de la unitat
RIGHT	Al·lòfon de la dreta de la unitat
L.TYPE	Tipus d'al·lòfon esquerra (consonant, vocal, ...)
L.PART	Lloc d'articulació de l'al·lòfon esquerra (oclusiu, palatal, ...)
L.MART	Mode d'articulació de l'al·lòfon esquerra (alveolar, fricatiu, ...)
L.SONOR	Sonoritat d'articulació de l'al·lòfon esquerra (sonor, sord, ...)
R.TYPE	Tipus d'al·lòfon dret (consonant, vocal, ...)
R.PART	Lloc d'articulació de l'al·lòfon dret (oclusiu, palatal, ...)
R.MART	Mode d'articulació de l'al·lòfon dret (alveolar, fricatiu, ...)
R.SONOR	Sonoritat d'articulació de l'al·lòfon dret (sonor, sord, ...)
PosInEG.L	Posició de l'al·lòfon esquerra en el grup d'entonació (Inicial, Mig, Final)
PosInEG.R	Posició de l'al·lòfon dret en el grup d'entonació (Inicial, Mig, Final)
PosInSyll.L	Posició de l'al·lòfon esquerra en la síl·laba (Inicial, Mig, Final)
PosInSyll.R	Posició de l'al·lòfon dret en el grup síl·laba (Inicial, Mig, Final)
PosInWord.L	Posició de l'al·lòfon esquerra en la paraula (Inicial, Mig, Final)
PosInWord.R	Posició de l'al·lòfon dret en la paraula (Inicial, Mig, Final)
PREV	Al·lòfon anterior a la unitat
NEXT	Al·lòfon posterior a la unitat
POS	Categoria gramatical
isSylStressed	Accentuació de la síl·laba
PREV.POS	Categoria gramatical de l'al·lòfon anterior
NEXT.POS	Categoria gramatical de l'al·lòfon posterior
PREV.TYPE	Tipus de l'al·lòfon anterior a la unitat (consonant, vocal, ...)
PREV.PART	Lloc d'articulació de l'al·lòfon anterior a la unitat (oclusiu, palatal, ...)
PREV.MART	Mode d'articulació de l'al·lòfon anterior a la unitat (alveolar, fricatiu, ...)
PREV.SONOR	Sonoritat de l'al·lòfon anterior a la unitat (sonor, sord, ...)
NEXT.TYPE	Tipus de l'al·lòfon posterior a la unitat (consonant, vocal, ...)
NEXT.PART	Lloc d'articulació de l'al·lòfon posterior a la unitat (oclusiu, palatal, ...)
NEXT.MART	Mode d'articulació de l'al·lòfon posterior a la unitat (alveolar, fricatiu, ...)
NEXT.SONOR	Sonoritat de l'al·lòfon posterior a la unitat (sonor, sord, ...)

Taula 5.12: Informació emprada per contextualitzar la subunitat (en cursiva, paràmetres ja emprats per caracteritzar la unitat).

En aquest canvi de nivell d'ajust, cal fer una distinció clara entre subcostos lingüístics

i context lingüístic per agrupar els pesos. Ambdós temes estan relacionats però s'empren per qüestions diferents: tal com s'ha explicat en l'apartat apartat 5.3, els subcostos de naturalesa lingüística mesuren la diferència que hi ha entre l'especificació lingüística d'una unitat i les seves possibles candidates dins la funció de cost. En canvi, per determinar els pesos que ponderen aquesta diferència a nivell de subunitat, s'empra informació lingüística del context, a part d'informació fonètica.

En aquesta contextualització lingüística per determinar quins pesos guien la selecció també s'empra la informació fonètica explicada en l'apartat 4.4. Així doncs, s'empren 20 paràmetres fonètics i lingüístics els quals es detallen en la taula 5.12.

Introduir els canvis que comporta aquest nou escenari en l'algorisme d'ajust automàtic mitjançant MLR/NNLS resulta relativament senzill tal i com es pot veure en els algorismes 5.1 (antic escenari) i 5.2 (nou escenari). Concretament, el canvi consisteix en incloure els passos 12 i 13 de l'algorisme 5.1 a dins del bucle que processa les versions (passos 8 i 9 en el nou algorisme). En les modificacions realitzades, s'assumeix que cada realització (versió) d'una unitat es correspon a un context diferent de la unitat dins la frase.

Algorisme 5.1 Algorisme d'ajust de pesos per totes les unitats del corpus mitjançant MLR/NNLS a nivell d'unitat.

procedure mlrTuning($\mathcal{U}$)

1: $W = \emptyset$
2: **for** $u \in \mathcal{U}$ **do**
3: $D = \emptyset$; $SC = \emptyset$
4: **for** $v \in versions(u)$ **do**
5: $\mathcal{W} = \{w : \forall_{w \in \mathcal{W}}(w \in versions(u) \wedge w \neq v)\}$
6: Es crea el conjunt $\mathcal{B}$ amb les 20 versions cepstralment més properes a la versió v.
$\mathcal{B} = \left\{b_{1,2,\ldots,20} : \forall_{b_i \in \mathcal{B}} \forall_{w \in \mathcal{W}}(b \in \mathcal{W} \wedge w \notin \mathcal{B} \wedge dist(b,v) < dist(w,v))\right\}$
7: S'assignen les distàncies de B a D_v
$D_v = \left\{d_i : \forall_{d_i \in D_v} \forall_{b_i \in \mathcal{B}}(d_i = dist(b_i, v))\right\}$
8: S'assignen els subcostos de B a SC_v
$SC_v = \left\{sc_{i,1\ldots M} : \forall_{sc_{i,1\ldots M} \in SC_v} \forall_{b_i \in \mathcal{B}}(sc_{i,j} = SC_i^j(b_i, v)) | j = \{PIT.T, \ldots, MFC.C\}\right\}$
9: $D = D \cup D_v$
10: $SC = SC \cup SC_v$
11: **end for**
12: Es calculen els pesos segons NNLS (algorisme 2.2) explicat en l'apartat 2.4.2
$w_u = NNLS(SC, D)$
13: $W = W \cup w_u$
14: **end for**
15: retorna W

Aquest canvi de nivell d'ajust també comporta canvis en l'entorn del GA. El canvi en

Algorisme 5.2 Algorisme d'ajust de pesos per totes les unitats contextualitzades del corpus mitjançant MLR/NNLS a nivell de subunitat.

procedure mlrTuning(U)

1: $W = \emptyset$
2: **for** $u \in \mathcal{U}$ **do**
3: **for** $v \in versions(u)$ **do**
4: $\mathcal{W} = \{w : \forall_{w \in \mathcal{W}}(w \in versions(u) \wedge w \neq v)\}$
5: Es crea el conjunt $\mathcal{B}$ amb les 20 versions cepstralment més properes a la versió v.
$\mathcal{B} = \left\{b_{1,2,\ldots,20} : \forall_{b_i \in \mathcal{B}} \forall_{w \in \mathcal{W}}(b \in \mathcal{W} \wedge w \notin \mathcal{B} \wedge dist(b,v) < dist(w,v))\right\}$
6: S'assignen les distàncies de B a D
$D = \left\{d_i : \forall_{d_i \in D_v} \forall_{b_i \in \mathcal{B}}(d_i = dist(b_i, v))\right\}$
7: S'assignen els subcostos de B a SC
$SC = \left\{sc_{i,1\ldots M} : \forall_{sc_{i,1\ldots M} \in SC_v} \forall_{b_i \in \mathcal{B}}(sc_{i,j} = SC_i^j(b_i, v)) | j = \{PIT.T, \ldots, Stress.L\}\right\}$
8: Es calculen els pesos segons NNLS (algorisme 2.2) explicat en l'apartat 2.4.2
$w_v = NNLS(SC, D)$
9: $W = W \cup w_v$
10: **end for**
11: **end for**
12: retorna W

l'entorn del GA resulta més rellevant ja que no hi havia una preselecció d'unitat de *target* a priori ni tampoc una preselecció de les seves 20 unitats més properes a l'hora de calcular els seus subcostos associats. Llavors, a nivell de subunitat ja no es pot realitzar el mostreig aleatori (pas 1 de l'algorisme 5.3) a l'hora de calcular la funció de *fitness* ja que aquest mostreig trencaria el requeriment de calcular els pesos només per la subunitat contextualitzada. És a dir, si a cada avaluació de pesos la versió que fa de *target* fos aleatòria, alhora es canviaria el context dels pesos avaluats i per tant els pesos no serien vàlids per un context fonètic/lingüístic específic.

En els algorismes 5.3 i 5.4 es mostra la versió l'ajust de pesos mitjançant GA a nivell d'unitat i en els algorismes 5.5 i 5.6 es mostra modificada per funcionar amb subunitats. El nou nivell de càlcul per versió contextualitzada es pot observar en el pas 3 de l'algorisme 5.6. Tanmateix, per minimitzar l'esmentat problema es redueix el nombre de versions més properes a la unitat desitjada de 20 a 5, en assumir que qualsevol d'elles, proporcionaria una millor síntesi. Aquest canvi es pot veure com el pas 3 de l'algorisme 5.3 es tradueix en el pas 2 de l'algorisme 5.5.

Algorisme 5.3 Algorisme pel càlcul del *fitness* en l'ajust de pesos per totes les unitats del corpus GA a nivell d'unitat.

procedure fitness(W,u)

1: v=random(versions(u))
2: $\mathcal{W} = \{w : \forall_{w \in \mathcal{W}}(w \in versions(u) \wedge w \neq v)\}$
3: Es crea el conjunt $\mathcal{B}$ amb les 20 versions cepstralment més properes a la versió v.
$\mathcal{B} = \left\{b_{1,2,\ldots,20} : \forall_{b_i \in \mathcal{B}} \forall_{w \in \mathcal{W}}(b \in \mathcal{W} \wedge w \notin \mathcal{B} \wedge dist(b,v) < dist(w,v))\right\}$
4: S'assignen els subcostos de B a SC
$SC = \left\{sc_{i,1\ldots M} : \forall_{sc_{i,1\ldots M} \in SC_v} \forall_{b_i \in \mathcal{B}}(sc_{i,j} = SC_i^j(b_i,v)) | j = \{PIT.T, \ldots, MFC.C\}\right\}$
5: El *fitness* és el valor dels subcostos ponderats ($SC * W^T$) de les unitats en $\mathcal{B}$
$F \leftarrow mean(SC * W^T)$
6: return F

Algorisme 5.4 Algorisme d'ajust de pesos per totes les unitats del corpus GA a nivell d'unitat.

procedure gaTuning(U, $SZ_{POP} = 200$, $IT = 1000$, $P_C = 0.6$, $P_M = 0.01$)

1: $W = \emptyset$
2: **for** $u \in \mathcal{U}$ **do**
3: $W_{u,0} = \{w_1, w_2, \ldots, w_{SZ_{POP}} | w_i = random()\}$
4: **for** $it = 1$ to $IT - 1$ **do**
5: $W_{u,it+1} \leftarrow$ creuament($W_{u,it+1}$,P_C) {Creuament unipunt explicat a l'apartat 3.2.3}
6: $W_{u,it+1} \leftarrow$ mutació($W_{u,it+1}$,P_M) {Mutació explicada a l'apartat 3.2.3}
7: $F = \left\{f_i : \forall_{f_i \in F}(f_i = fitness(w_i,u) \wedge w_i \in W_{u,it})\right\}$ {Càlcul de *fitness* detallat en l'algorisme 5.3}
8: $W_{u,it+1} \leftarrow$ tournament($W_{u,it}$,F) {Selecció per torneig explicada a l'apartat 3.2.3}
9: **end for**
10: $w_u \leftarrow mean(W_{u,1}, W_{u,2}, \ldots, W_{u,IT})$
11: $W \leftarrow W \cup w_u$
12: **end for**
13: retorna W

Algorisme 5.5 Algorisme pel càlcul del *fitness* en l'ajust de pesos per totes les unitats del corpus GA a nivell de subunitat.

procedure fitness(W,u,v)

1: $\mathcal{W} = \{w : \forall_{w \in \mathcal{W}}(w \in versions(u) \wedge w \neq v)\}$
2: Es crea el conjunt $\mathcal{B}$ amb les 5 versions cepstralment més properes a la versió v.
$\mathcal{B} = \left\{b_{1,2,\ldots,5} : \forall_{b_i \in \mathcal{B}} \forall_{w \in \mathcal{W}}(b \in \mathcal{W} \wedge w \notin \mathcal{B} \wedge dist(b,v) < dist(w,v))\right\}$
3: S'assignen els subcostos de B a SC
$SC = \left\{sc_{i,1\ldots M} : \forall_{sc_{i,1\ldots M} \in SC_v} \forall_{b_i \in \mathcal{B}}(sc_{i,j} = SC_i^j(b_i,v)) | j = \{PIT.T, \ldots, Stress.L\}\right\}$
4: El *fitness* és el valor dels subcostos ponderats ($SC * W^T$) de les unitats en $\mathcal{B}$
$F = mean(SC * W^T)$
5: return F

Algorisme 5.6 Algorisme d'ajust de pesos per totes les unitats del corpus GA a nivell de subunitat.

procedure gaTuning(U, $SZ_{POP} = 200$, $IT = 1000$, $P_C = 0.6$, $P_M = 0.01$)

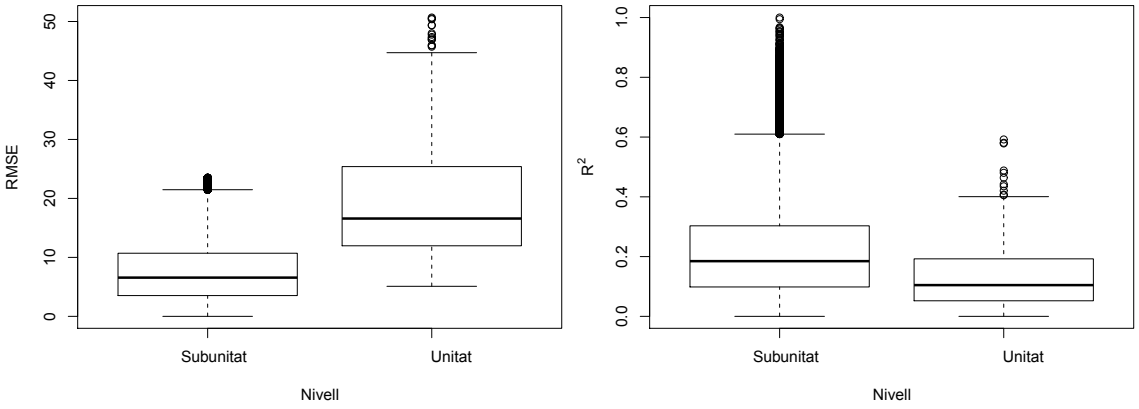
1: $W = \emptyset$
2: **for** $u \in \mathcal{U}$ **do**
3: **for** $v \in versions(u)$ **do**
4: $W_{v,0} = \{w_1, w_2, \ldots, w_{SZ_{POP}} | w_i = random()\}$
5: **for** $it = 1$ to $IT - 1$ **do**
6: $W_{u,it+1} \leftarrow$ creuament($W_{u,it+1}$,P_C) {veure apartat 3.2.3}
7: $W_{u,it+1} \leftarrow$ mutació($W_{u,it+1}$,P_M) {veure apartat 3.2.3}
8: $F = \left\{f_i : \forall_{f_i \in F}(f_i = fitness(w_i, u, v) \wedge w_i \in W_{u,it})\right\}$ {Càlcul de *fitness* detallat en l'algorisme 5.5}
9: $W_{u,it+1} \leftarrow$ tournament($W_{u,it}$,u,v,F) {veure l'apartat 3.2.3}
10: **end for**
11: $w_v = mean(W_{v,1}, W_{v,2}, \ldots, W_{v,IT})$
12: $W = W \cup w_v$
13: **end for**
14: **end for**
15: retorna W

5.4.3 Comparació de resultats: unitat vs. subunitat

Un cop s'han exposat els canvis que comporta treballar a nivell de subunitat, s'estudia si aquest pas comporta un augment de la precisió (disminució de l'error) dels pesos obtinguts segons l'ajust automàtic i conseqüentment això condueix a una millora en la fiabilitat dels mateixos segons les mètriques RMSE i R^2 (per l'aproximació segons regressió lineal) i una disminució de la desviació típica sobre la mitjana (per l'aproximació segons GA).

Millora de la consistència en la regressió lineal

La figura 5.13 mostra les mètriques de fiabilitat R^2 i RMSE per l'ajust dels pesos mitjançant regressió lineal del corpus *uvig_dav_es* segons si es treballa a nivell d'unitat o de subunitat. Aquests resultats es detallen a la taula 5.13. Tal i com es pot observar, treballar a nivell de subunitat comporta una millora significativa en ambdós índex de consistència. Concretament, l'error RMSE (l'error de predicció) passa de 23.98 ± 25.9 a 9.307 ± 11.27 a nivell de subunitat. Pel que fa al coeficient de determinació R^2 es passa de modelar un $33.9 \pm 14.36\%$ de les dades a un $43.4 \pm 17.42\%$. En ambdues comparatives s'obté una significància de $p < 2 \cdot 10^{-16}$ en aplicar la prova t de Student per parelles (Hochberg, 1988). Val a dir que l'error 0 o el coeficient de determinació 0 s'obté en el cas de realitzar la regressió



(a) RMSE obtingut en aplicar MLR/NNLS per unitat i subunitat

(b) R^2 obtingut en aplicar MLR/NNLS per unitat i subunitat

Figura 5.13: Mètriques de fiabilitat obtingudes quan s'aplica la regressió lineal per obtenir els pesos a nivell d'unitat i subunitat contextualitzada *uvig_dav_es*.

amb poques dades.

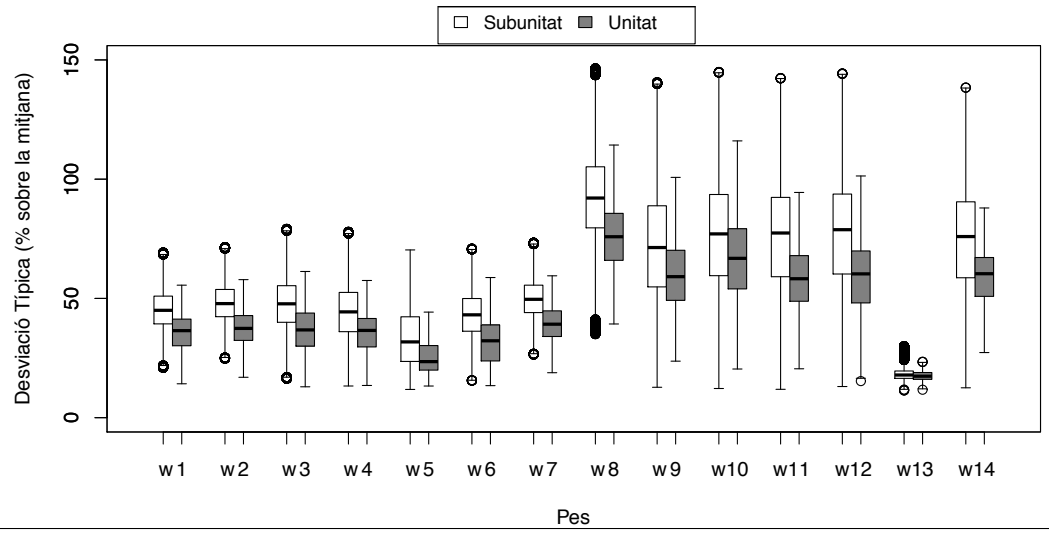
Estadística	Nivell	Min	1Q	Mediana	Mitjana	3Q	Max	σ
RMSE	Unitat	5.094	12.138	17.205	23.897	18.106	314.486	25.902
	Subunitat	0.000	3.673	6.906	9.307	11.605	480.202	11.27
R^2	Unitat	0.2%	22.8%	32.3%	33.9%	43.9%	76.9%	14.36%
	Subunitat	0%	31.4%	43%	43.4%	55%	100%	17.42%

Taula 5.13: Mètriques de fiabilitat obtingudes quan s'aplica la regressió lineal per obtenir els pesos a nivell d'unitat i subunitat contextualitzada *uvig_dav_es*.

Efectes en l'ajust mitjançant GA

Tal com s'ha exposat a l'apartat 4.3.2 per analitzar la consistència dels pesos s'avalua com a mesura de consistència el % de variació (desviació típica) de valor del pes respecte el seu valor mig.

En la figura 5.14 es pot veure detallat per pes i nivell d'ajust el percentatge de variació del pes (desviació típica sobre mitjana) segons s'hagi fet l'ajust de pesos per unitat o subunitat. En aquest cas el fet de treballar a nivell de subunitat empitjora l'estadístic de desviació típica emprat. Concretament es passa d'una desviació típica mitjana de $45.5 \pm 17.06\%$



Subcost	Nivell	Min	1Q	Mediana	Mitjana	3Q	Max
w_1 =PIT.T	Subunitat	20.96%	39.33%	45.01%	45.11%	50.96%	69.24%
	Unitat	14.21%	30.15%	36.49%	35.81%	41.33%	55.56%
w_2 =ENE.T	Subunitat	24.74%	42.34%	47.82%	48.21%	53.76%	71.54%
	Unitat	16.93%	32.40%	37.43%	37.27%	42.81%	57.88%
w_3 =DURL.T	Subunitat	16.35%	39.99%	47.74%	47.31%	55.34%	79.14%
	Unitat	12.95%	30.01%	36.79%	36.57%	43.86%	61.29%
w_4 =DURR.T	Subunitat	13.28%	36.06%	44.36%	44.19%	52.52%	77.90%
	Unitat	13.51%	29.66%	36.58%	35.93%	41.58%	57.52%
w_5 =PIT.C	Subunitat	11.84%	23.58%	31.77%	33.52%	42.30%	70.33%
	Unitat	13.27%	19.94%	23.50%	25.56%	30.23%	44.26%
w_6 =ENE.C	Subunitat	15.54%	36.25%	43.12%	42.98%	49.95%	70.87%
	Unitat	13.43%	23.78%	32.23%	31.64%	38.88%	58.76%
w_7 =MFC.C	Subunitat	26.44%	44.06%	49.63%	49.88%	55.56%	73.37%
	Unitat	18.85%	34.03%	39.19%	39.38%	44.79%	59.49%
w_8 =PosInEG.L	Subunitat	35.04%	79.59%	92.09%	92.21%	105.19%	146.48%
	Unitat	39.28%	65.96%	75.87%	75.26%	85.67%	114.31%
w_9 =PosInSyl.L	Subunitat	12.74%	54.83%	71.33%	71.65%	88.85%	140.66%
	Unitat	23.70%	49.20%	59.12%	60.16%	70.19%	100.75%
w_{10} =PosInWord.L	Subunitat	12.23%	59.54%	77.05%	76.09%	93.57%	144.95%
	Unitat	20.41%	53.99%	66.80%	66.46%	79.22%	116.05%
w_{11} =Prev.L	Subunitat	11.89%	59.10%	77.43%	75.04%	92.36%	142.41%
	Unitat	20.50%	48.80%	58.26%	57.97%	67.94%	94.43%
w_{12} =Next.L	Subunitat	13.06%	60.24%	78.82%	76.28%	93.76%	144.30%
	Unitat	15.34%	48.11%	60.29%	58.47%	69.91%	101.34%
w_{13} =POS.L	Subunitat	11.40%	16.49%	17.89%	18.25%	19.58%	29.94%
	Unitat	11.72%	16.06%	17.39%	17.53%	18.93%	23.41%
w_{14} =Stress.L	Subunitat	12.52%	58.66%	75.93%	73.97%	90.50%	138.39%
	Unitat	27.30%	50.86%	60.36%	58.87%	67.16%	87.93%

Figura 5.14: Desviacions típiques (en % sobre el valor de la mitjana) del valor dels pesos en el transcurs de 500 generacions quan s'aplica GA a nivell d'unitat o subunitat.

a $56.7 \pm 20.7\%$. Si s'aplica la prova t de Student per parelles amb correcció de Bonferroni (Hochberg, 1988) es conclou que $STD_{Unitat} < STD_{Subunitat}$ amb una significança de $p < 1.6 \cdot 10^{-9}$ per a tots els pesos.

En aquest cas es pot concloure que el fet de treballar a nivell de subunitat no comporta una millora en la consistència dels pesos obtinguts quan l'ajust de pesos es realitza mitjançant algorismes genètics. Per explicar aquest resultat sorgeixen dues possibles hipòtesis: *i*) la pèrdua de l'aleatorietat en la funció de *fitness* implementada cosa que implica un canvi important en les assumpcions del disseny original (el GA s'havia dissenyat ad hoc per un entorn sorollós (Alías i Llorà, 2003) tal com s'ha explicat en l'apartat 3.2.3), i *ii*) s'ha assumit, sense contrastar-ho, que la funció de cost ponderada obtinguda en les 5 versions cepstralment més properes s'hi troba el *fitness* a minimitzar de manera unívoca (sense considerar la possibilitat que l'amitjanat no considera l'ordre entre elles). Llavors, seria més apropiat avaluar el vector de pesos penalitzant els canvis de posició que infereixen els pesos respecte l'ordre de candidates obtingut a través de distàncies cepstrals.

Com que el GA emprat ha estat simplement una adaptació del treball de (Alías i Llorà, 2003), el qual era només una prova de viabilitat, i tenint en compte que el seu algorisme fou dissenyat específicament per ajustar els pesos a nivell d'unitat i soroll, en aquesta tesi doctoral només es manifesta el problema observat i es deixa la recerca en l'ajust de pesos automàtic mitjançant computació evolutiva per a posteriors estudis que haurien de dissenyar una millor funció de *fitness*. Aquesta motivació també ve motivada perquè la regressió MLR/NNLS gaudeix d'una millora considerable de la fiabilitat.

Finalment, cal afegir que la naturalesa discreta dels pesos lingüístics comporta que aquests presentin un índex de desviació típica més alt que els pesos acústics, amb excepció del pes de categoria gramatical POS.L. En aquest sentit es pot remarcar que el pes de categoria gramatical (POS.L) presenta un comportament diferenciat respecte els pesos acústics i lingüístics en l'ajust evolutiu.

5.4.4 Elecció del mètode automàtic per detectar patrons

Després d'aplicar l'ajust de pesos a nivell de subunitat es pot confirmar que aquest canvi de nivell de detall provoca una millora substancial en la robustesa el modelat del problema quan s'apliquen tècniques de regressió lineal clàssiques però en canvi provoca un empitjorament si s'empren les tècniques d'ajust evolutives proposades per (Alías i Llorà, 2003).

No obstant això, al treballar a nivell de subunitat contextualitzada el GA empitjora

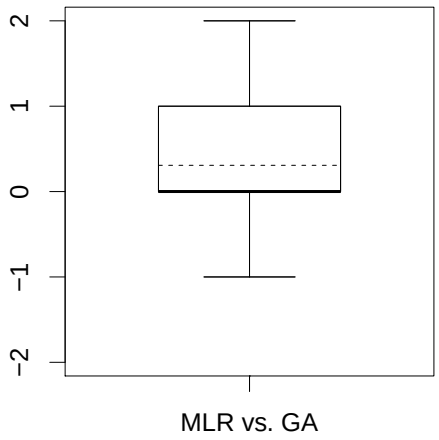


Figura 5.15: CMOS dels pesos obtinguts mitjançant tècniques d'ajust automàtic.

la seva fiabilitat. És de rigor destacar que el disseny original del GA (Alías i Llorà, 2003) assumeix que amitjanant el cost total obtingut amb la ponderació segons els pesos avaluats dels subcostos de les 5 versions cepstralment més properes, s'obté una avaluació dels pesos de manera unívoca. Però en canvi el disseny no té en compte que l'amitjanat no considera l'ordre entre elles. A més, el disseny del GA obeeix a una entorn sorollós seleccionant aleatòriament unitats diferent per avaluar cada combinació de pesos. Aquesta premissa no s'ha respectat al treballar a nivell de subunitat. Per tant, el motiu de l'empitjorament del GA ve més donat per un mal disseny que no pas per un nivell incorrecte d'ajust.

Convé destacar que amb l'objectiu d'obtenir una avaluació preliminar entre els dos mètodes d'ajust automàtic, amb anterioritat a aquest capítol es va realitzar una prova perceptiva en que els diferents usuaris comparessin la qualitat de diferents frases sintetitzades segons els diferents pesos (MLR/GA) a nivell de subunitat. No obstant, al ser preliminar, aquesta prova perceptiva es va dur a terme amb un corpus diferent (*url_pat_es*) de l'utilitzat en aquest capítol (*uvig_dav_es*). El corpus, propi del grup de recerca, és inferior en termes d'extensió (42 min. en comptes de 1.9h.). Resumidament el corpus *url_pat_es* conté 833 locucions (en comptes de 1217) formades per 5468 paraules (en comptes de 17797) que proporcionen cobertura per 1709 paraules diferents (en comptes de 5465). A nivell d'unitats el corpus té un total de 27153 unitats (en comptes de 88571) que es divideixen en 763 unitats diferents (en comptes de 827).

Es van escollir del corpus 30 frases diferents de manera aleatòria, que varen ser avaluades per 13 usuaris mitjançant un MOS Comparatiu (CMOS). Els resultats de la comparació es mostren a la figura 5.15 on es mostra una clara preferència dels usuaris pels pesos ajus-

tats mitjançant MLR. Analíticament MLR>GA de manera significativa ($p = 0.0348$) segons la prova de test de Wilcoxon (Hollander i Wolfe, 1973).

Després dels resultats de fiabilitat obtinguts i observant els resultats de la prova preliminar, s'escullen els pesos obtinguts mitjançant regressió lineal com a conjunt de dades útils per cercar-hi patrons de comportament mitjançant *clustering* i així definir els clústers on realitzar l'ajust interactiu.

5.5 Millora de la metodologia de *clustering* dels pesos

5.5.1 Consideracions prèvies

En l'apartat 4.4 s'han exposat les avantatges de realitzar l'ajust dels pesos de la funció de cost a nivell de grups d'unitats o clústers. Poder agrupar els pesos calculats prèviament de manera automàtica permet obtenir un ajust interactiu dels diferents patrons de pesos trobats a partir de l'optimització automàtica. Llavors resulta factible la participació humana al no tenir els inconvenients de la fatiga o contradiccions de l'usuari.

Tal com s'ha esmentat en l'apartat 5.4.1, el fet de treballar a nivell de subunitat no comporta canvis substancials en l'agrupament d'aquests pesos ja que simplement s'incorpora més informació per realitzar l'agrupament. Aquesta informació és la que fa referència als diferents contextos acústics i fonètics de cada vector de pesos (taula 5.12).

Per tal de realitzar el *clustering* del capítol anterior s'emprava la implementació *wagon* (Black i Taylor, 1997*a*) de l'algorisme CART (Breiman *et al.*, 1984) per definir els grups d'ajust. La tria de l'algorisme es va basar principalment en criteris històrics ja que és un algorisme àmpliament emprat en l'àmbit de la síntesi per selecció d'unitats (Black i Taylor, 1997*a*; Kominek i Black, 2005). Tanmateix, per a realitzar el procés de *clustering* els autors identifiquen en l'algorisme CART dues mancances que s'exposen a continuació (Kominek i Black, 2005):

i) *Clustering predictiu:* El fenomen del *clustering* predictiu apareix en combinar l'agrupament mitjançant una mètrica de distància amb un conjunt de predictors - en aquest cas, les característiques esmentades en la taula 5.12 -. Aquests predictors s'empren per trobar els punts de divisió (organitzats segons un arbre de decisió binari) que millor modelen les diferències entre els grups (en aquest cas grups de pesos). Aquest fet comporta que l'agrupament depengui dels valors que prenguin el conjunt de predictors considerats. Conseqüentment, si una separació natural dels pesos no es pot inferir

a partir de les característiques lingüístiques i fonètiques que conformen el conjunt de predictors s'assumeix que aquesta separació no existeix.

ii) *Arbres de decisió equilibrats*: El *wagon* permet generar arbres de decisió binaris de manera equilibrada tot i que no sempre coincideixi amb la distribució de les dades d'entrada. A partir d'unes proves realitzades (Kominek i Black, 2005) sobre el corpus *rms_arctic* (Kominek i Black, 2004) es conclou que obtenir clústers amb un nombre semblant d'elements pot comportar un alt error de predicció (en el seu cas un 47% d'error quan el màxim error en classificació binària és del 50%)

5.5.2 Fases de l'agrupament: *Clustering* i classificació

En aquest apartat es descriu una aproximació al problema de la detecció de patrons que obeeix millor a una metodologia clàssica dins l'àmbit de l'aprenentatge artificial. Aquesta nova aproximació permet superar els problemes del *clustering* predictiu i les restriccions dels arbres equilibrats. El problema de la detecció de patrons de pesos inferits per la informació lingüística i fonètica es pot dividir en dos subproblemes amb objectius diferents: *i*) detecció independent dels grups de dades (aprenentatge no supervisat), i *ii*) modelat dels grups trobats de manera natural a partir del conjunt de predictors associat. En altres paraules, el problema es subdivideix (veure figura 5.16) en un un problema clàssic de *clustering* enfocat a trobar els patrons de pesos naturals (no coneguts) i un problema de classificació que donat un context fonètic i lingüístic d'una determinada unitat associa el vector (patró) de pesos a utilitzar en cada cas.

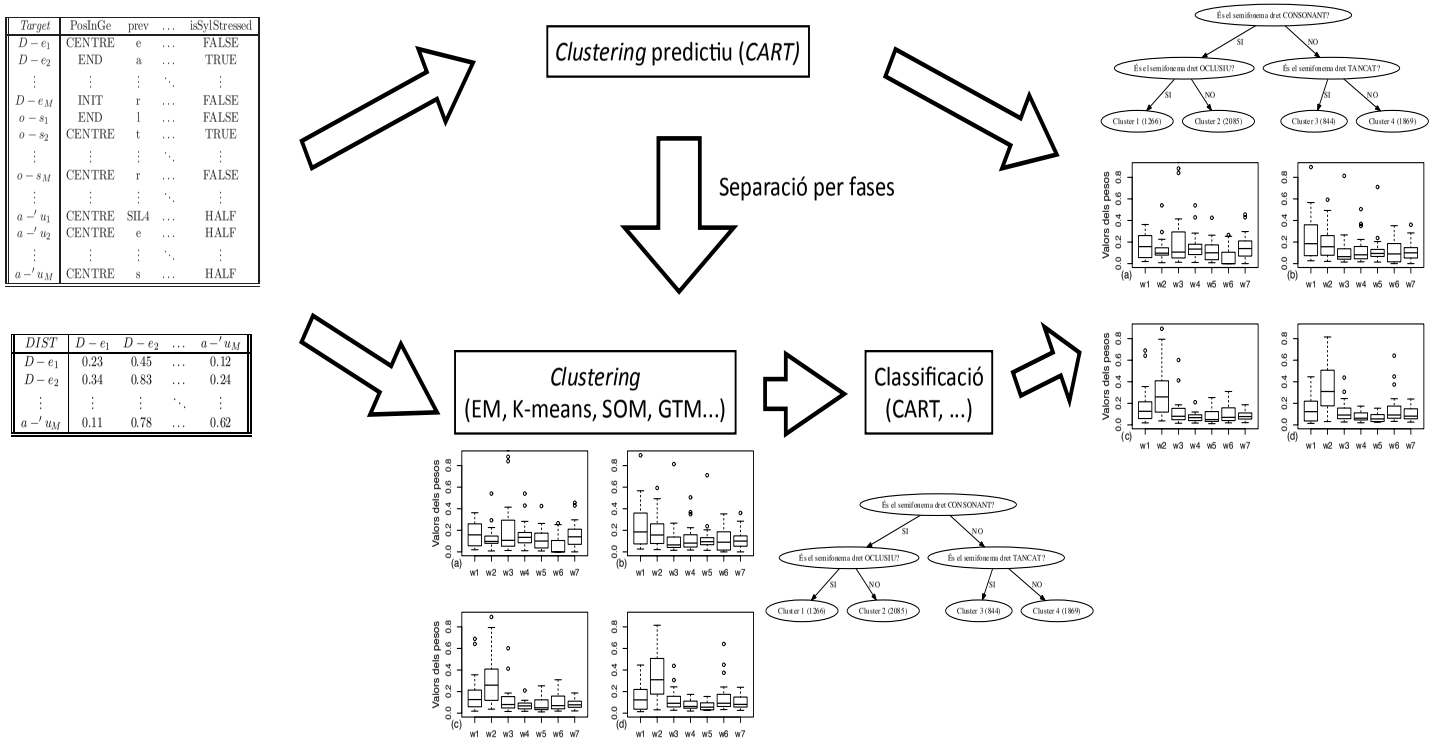


Figura 5.16: Separació del problema del *clustering* predictiu en dues fases: detecció de patrons (*clustering*) i classificació.

5.5.3 Comparativa d'algorismes de *clustering*

L'objectiu principal de l'etapa de *clustering* és trobar aquells grups de pesos que conformen patrons de manera natural amb independència dels contextos lingüístics i fonètics (grups de predictors) associats. En la literatura es proposen diferents algorismes per aquesta tasca on cadascun d'ells té els seus avantatges i inconvenients. En el treball de (Jain *et al.*, 1999) es realitza una revisió exhaustiva de les diferents metodologies de *clustering* juntament amb la seva naturalesa. Principalment els mètodes de *clustering* es divideixen entre mètodes basats en partició i mètodes basats en jerarquia (Jain *et al.*, 1999). La diferència entre ambdós tipus de mètodes arrela en la dependència del càlcul d'un nombre determinat de clústers. Si per dividir les dades en k grups s'ha de realitzar prèviament el càlcul de $k-1$ (divisiu) o $k+1$ (aglomeratiu) la metodologia de *clustering* és jeràrquica. En canvi, si els k grups es poden inferir directament a partir de les dades d'origen, llavors es parla d'algorismes basats en partició (Jain *et al.*, 1999).

L'algorisme *wagon* emprat fins el moment, s'ubicaria dins del grup d'algorismes basats en jerarquia. Altres classificacions dels algorismes de *clustering* inclouen l'aleatorietat del seu comportament - determinista si retorna sempre els mateixos grups per les mateixes dades d'entrada o estocàstic en cas contrari - o bé el tipus de sortida que dóna - mètodes forts si assigna cada instància del conjunt d'entrada a un grup o mètodes difusos si torna el respectiu grau de pertinença de cada instància a cada grup.

Algorismes de *clustering*

L'algorisme emprat fins el moment, el *wagon*, es tipificaria com un algorisme jeràrquic, fort i determinista. No obstant, a priori no se sap quina és la metodologia de *clustering* idònia per cercar els patrons de pesos que millor representin els comportaments que s'han trobat en el *corpus*. Per aquest motiu es seleccionen 6 algorismes de referència de l'estat de l'art i es fan competir entre ells per a seleccionar-ne el millor segons les mètriques explicades en l'apartat 4.4.4, les quals serveixen per determinar la bondat dels diferents grups trobats. Els mètodes de *clustering* es descriuen breument a continuació:

- *Agglomerative Hierarchical (H)* (Johnson, 1967): Aquest mètode considera en primer lloc cadascuna de les instàncies d'entrada com un sol grup i les va fusionant progressivament fins assolir el nombre de clústers desitjat. El criteri de fusió normalment es pren considerant la distància més petita (entre dos vectors de pesos diferents) de la matriu de distàncies d'entrada.

- *Mapes autoorganitzatius (*Self-organizing Maps – *SOM)* (Kohonen, 1990): Els mapes autoorgantizatius basen el seu funcionament en el modelat de les dades d'entrada mitjançant una xarxa neuronal bicapa. L'objectiu del modelat és que les neurones especialitzades de la xarxa neuronal (variables latents) adoptin una resposta positiva d'acceptació (s'activin) en funció dels diferents patrons de pesos presentats a la xarxa. Llavors, els patrons de pesos que activen una determinada neurona formen un clúster. Aquest tipus d'aprenentatge s'anomena aprenentatge competitiu. Val a dir que el comportament dels mapes autoorganitzatius és estocàstic, no es poden garantir dos models idèntics per les mateixes dades d'entrada.

- *Mapes topogràfics generatius (*Generative Topographic Mapping – *GTM)* (Bishop *et al.*, 1998): L'aproximació dels mapes topogràfics generatius s'inspira en el model de variable latent dels mapes autoorganitzatius ja explicats. En els GTM, l'espai latent és una malla (d'una o dues dimensions) de gaussianes les quals s'assumeix com una projecció no lineal de l'espai d'entrada.

- *Expectation Maximization (EM)*(Fraley i Raftery, 1998): L'aproximació segons l'algorisme esperança-maximització (*expectation-maximization*) modela les dades a través de diferents gaussianes igual que el mètode GTM ja explicat. A diferència del GTM l'algorisme d'EM no considera les gaussianes com la projecció de les dades d'entrada en un espai latent en forma de malla sinó que les gaussianes no adopten cap restricció de veïnatge per modelar les esmentades dades. L'algorisme EM - de naturalesa iterativa - s'empra per ajustar els paràmetres (μ, σ) de les diverses gaussianes que es consideren per a modelar un espai de dades donat. Tanmateix l'algorisme té diferents finalitats més enllà del *clustering*, que inclouen la classificació i la regressió, entre d'altres (Moon, 1996). El procés de trobar els paràmetres del model es coneix en la literatura com a procés de màxima versemblança (*maximum likelihood*) si no s'empra cap distribució a-priori o màxim a posteriori (MAP) si altrament s'empra una distribució a-priori (McLachlan i Krishnan, 1997).

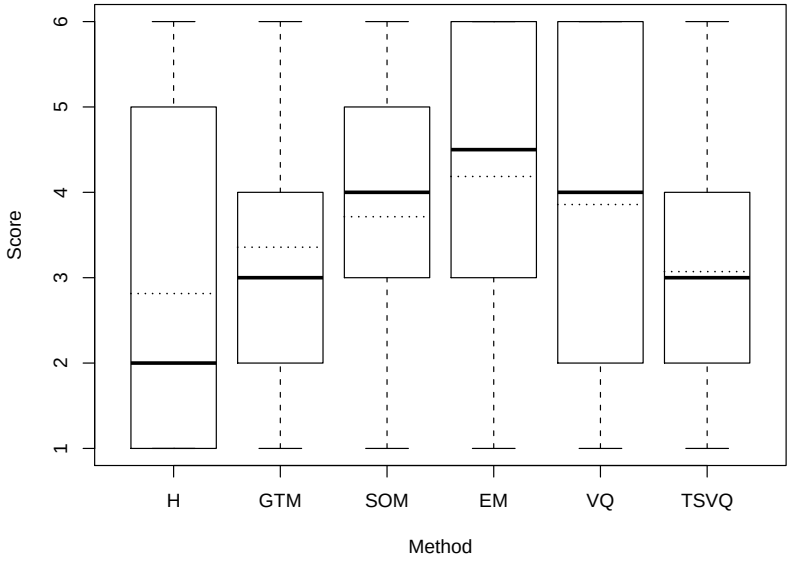
- *Quantificació vectorial (VQ o k-means)* (Hartigan i Wong, 1979): L'agrupament mitjançant quantificació vectorial (*k-means*) és un mètode que assigna les dades als grups en funció de la distància (en el nostre cas distància del cosinus – veure apartat 4.4.4) a una instància fictícia que representa els valors mitjans de les diferents instàncies que es troben en el clúster. Aquest vector fictici s'anomena centroide. En aquest sentit, la quantificació vectorial s'assimila a l'aproximació EM ja que intenta trobar els centroides dels clústers de manera iterativa. Tanmateix, a diferència de l'EM el comportament de la quantificació vectorial és estocàstic i per tant no es poden garantir dos

models idèntics per les mateixes dades d'entrada.

- *Quantificació vectorial estructurada en arbre (TSVQ)* (Gersho i Gray, 1992): L'agrupament mitjançant TSVQ és una variant de l'algorisme VQ tenint en compte la propietat que l'VQ es comporta millor, a nivell de solidesa i fiabilitat dels grups, quan realitza particions binàries. Llavors, l'algorisme TSVQ realitza particions binàries de manera recursiva fins a trobar el nombre de clústers especificat a priori, establint així una jerarquia entre les particions de manera divisiva.

Comportament dels algorismes

Un cop s'han recopilat els diferents algorismes de *clustering* típics de l'estat de l'art (Jain *et al.*, 1999), es compara la seva adequació al problema de la detecció de patrons en els pesos obtinguts mitjançant els mètodes d'ajust automàtics. A tal efecte s'usen les mètriques exposades en l'apartat 4.4.4 (veure taula 5.18), que permeten avaluar la resposta dels algorismes davant el problema plantejat. Com que les diferents mètriques no es poden comparar entre elles, la comparació es realitza anàlogament a la selecció per *ranking* explicada en l'apartat 3.2.3, mitjançant la qual a cada mètode se li assignen uns punts en funció de la posició que pren en la comparativa. Concretament, al comparar 6 mètodes diferents s'assignen 6 punts al millor mètode, 5 al segon, i així fins a arribar a 1 punt per al pitjor (sisè) mètode. Primer s'assigna la puntuació per cada dupla <mètrica, nombre de clústers (estudiant d'1 a 10)> i llavors es selecciona l'algorisme de *clustering* que ha obtingut millors puntuacions per totes les mètriques i nombre de grups analitzats. Es considera un topall màxim de 10 grups ja que un nombre superior de grups seria inassolible pels usuaris en termes de fatiga. Els resultats es mostren a la figura 5.17(a), on es pot observar que el millor mètode és l'EM seguit del VQ clàssic i el SOM, no havent-hi diferències significatives entre ells. Llavors, obeint els resultats obtinguts, s'escull l'algorisme EM per trobar els patrons de pesos existents en el corpus perquè és el que presenta un millor comportament a través de totes les mètriques i nombre de grups analitzat. A nivell de significància estadística (taula de la figura 5.17(b)) l'algorisme EM és el que més diferències (3) presenta respecte la resta de metodologies de *clustering*. Malgrat que els resultats d'EM no són significativament millors que SOM i VQ, aquests dos últims presenten menys diferències significatives (2) amb la resta d'algorismes. S'ha de tenir en compte que per aquestes proves no s'ha emprat la prova t de Student ja que les dades no segueixen una distribució normal (són dades discretes amb una clara asimetria). Seguint el treball de (Hollander i Wolfe, 1973) s'ha emprat la prova U de Mann-Whitney (Mann i Whitney, 1947) per a distribucions de dades no paramètriques, que permet obtenir la significança de la diferència entre dues distribucions



(a) Comparativa de la posició que adopta cada mètode per totes les duples <mètrica, nombre de clústers> analitzada. El mètode més ben posicionat adopta 6 punts i el pitjor adopta 1 punt (el valor de la mitjana es mostra en la línia de punts horitzontal de cada *boxplot*).

	EM	GTM	H	SOM	TSVQ
GTM	**0.0015**	-	-	-	-
H	$\mathbf{3 \cdot 10^{-4}}$	0.1313	-	-	-
SOM	0.1295	0.1532	**0.0112**	-	-
TSVQ	$\mathbf{8 \cdot 10^{-4}}$	0.3446	0.438	**0.0203**	-
VQ	0.3603	0.0799	**0.0065**	0.6403	**0.0168**

(b) Significances del les diferències entre els mètodes segons la prova U de *Mann-Whitney* (Mann i Whitney, 1947) (en negreta els valors estadísticament significatius).

Figura 5.17: Diferències i significances entre els diferents mètodes aplicats per trobar patrons de pesos (mitjançant ajust automàtic) en el corpus *uvig_dav_es*.

no gaussianes.

Tria del nombre de clústers

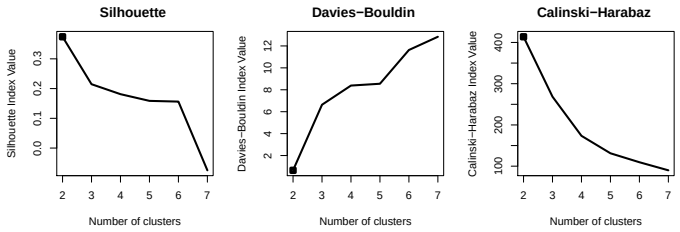
La tria del nombre de clústers es realitza de la mateixa manera que s'ha exposat en l'apartat 4.4.4, però canviant l'algorisme *wagon* per l'algorisme EM, pels motius que s'acaben de comentar. En aquest cas, pel fet de tenir un gran volum de dades s'avalua obtenir un nombre de clústers entre 2 i 8.

Els resultats es mostren a la figura 5.18(a). En aquesta taula, es pot observar que no hi ha consens entre les mètriques a l'hora d'escollir el nombre de clústers òptim. Si s'analitzen les gràfiques es poden distingir clarament dos grups de mètriques: el primer grup el conformen les mètriques Silhouette, Davies-Bouldin, Calinski-Harabaz, Dunn i Krzanowski-Lai. El segon grup el conformen les mètriques *C-Index*, Hartigan, *weighted inter/intra i Homogenity*. El fet diferencial de les mètriques del primer grup es troba totes elles empitjoren en augmentar el nombre de clústers, per tant són conservadores (escollint sempre el nombre mínim de clústers) al no veure una divisió molt clara. En l'entorn de l'ajust de pesos, per ser un entorn sorollós (Alías i Llorà, 2003), és normal que no es puguin trobar divisions evidents. No obstant, la mètrica Krzanowski-Lai té un segon màxim a 5 clústers en considerar una millora evident en les divisions $3 \rightarrow 4$ i $4 \rightarrow 5$. Aquesta última dada es correspon amb la majoria de les mètriques del segon grup, que determinen de forma majoritària que el nombre de clústers es troba entre 5 i 6 (amb excepció de la mètrica Hartigan, que opta per 3 clústers).

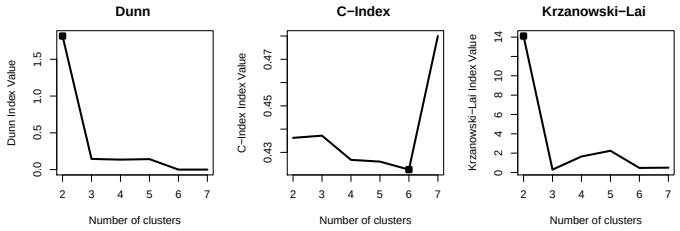
Tenint en compte els resultats obtinguts, la tria del nombre de clústers no resulta tant evident com ho ha estat en el capítol anterior (apartat 4.4.4). Pel fet de treballar en un entorn sorollós es descarten les mètriques conservadores (Silhouette, Davies-Bouldin, Calinski-Harabaz i Dunn) i, com a subcriteri, s'adopta el valor mig de la resta de mètriques. Per tant, el nombre de clústers escollit al final és 5 tenint en compte que un nombre superior de clústers requeriria un nombre excessiu d'usuaris i proves a ajustar. Aquest mateix subcriteri no hauria variat el nombre de clústers en l'agrupament realitzat en l'apartat 4.4.4. En aquest sentit, s'escull la mediana del nombre indicat per les mètriques no conservadores (*C-Index*,Hartigan,*weighted inter/intra*,*Homogeneity* i Krzanowski-Lai). Convé denotar que l'únic impacte en l'ajust perceptiu d'un nombre excessiu de grups és la confusió de les seves unitats, implicant redundància en els pesos ajustats (mateix patró de pesos per dos clústers diferents).

Indicador/#clusters	2	3	4	5	6	7
Silhouette	**0.3742**	0.2147	0.1812	0.1587	0.1562	-0.0748
Davies-Bouldin	**0.6502**	6.6352	8.384	8.5498	11.629	12.836
Calinski-Harabaz	**413.8**	268.5	173.57	131.24	109.56	89.944
Dunn	**1.816**	0.1449	0.1356	0.1425	0	0
C-Index	0.4362	0.4372	0.4268	0.426	**0.4226**	0.48
Krzanowski-Lai	**14.11**	0.2996	1.654	2.2463	0.4727	0.5003
Hartigan	92.17	**-11**	3.2711	16.23	-5.2732	-37.663
weighted inter/intra	0.688	0.7297	0.73	**0.7448**	0.7326	0.6999
Homogenity	-0.1434	-0.0237	0.0404	0.0488	**0.0583**	-0.0288

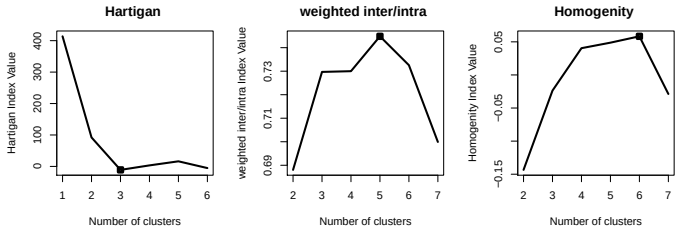
(a) Resultats de les mètriques de validació (Wang *et al.*, 2009) detallades per mètrica i nombre de clústers.



(b) Silhouette,Davies-Bouldin,Calinski-Harabaz.

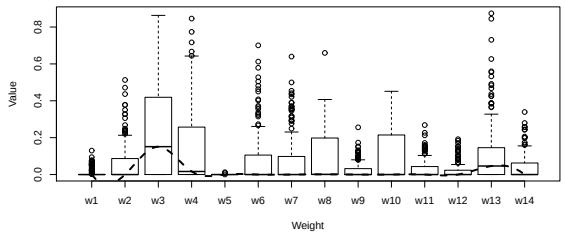


(c) Dunn,*C-Index*,Krzanowski-Lai.

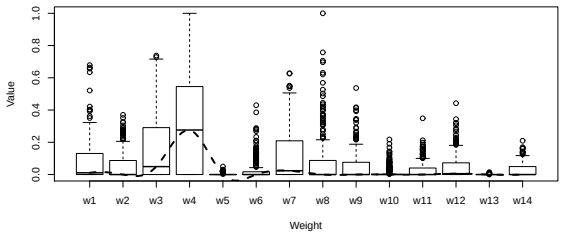


(d) Hartigan,*weighted inter/intra* i *Homogenity*.

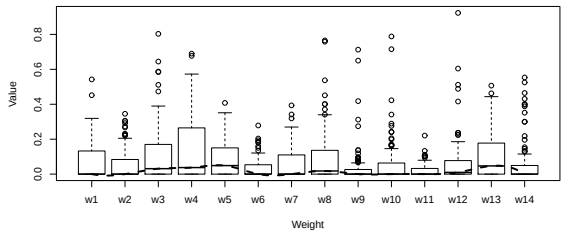
Figura 5.18: Indicadors obtinguts quan s'aplica EM als pesos obtinguts mitjançant ajust automàtic del corpus *uvig_dav_es* (el nombre de clústers òptim s'índica amb un quadre).



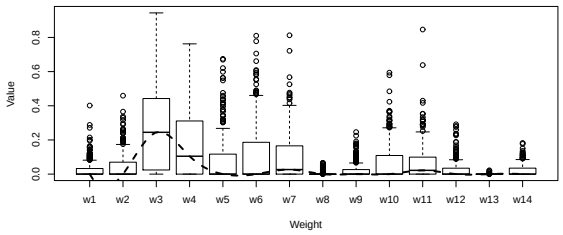
(a) 1r clúster (210 pesos).



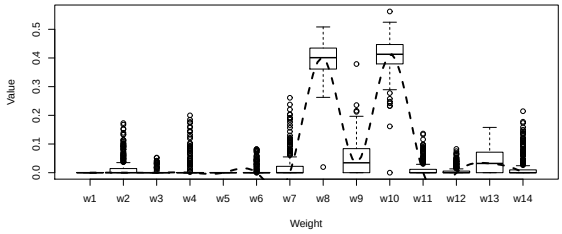
(b) 2n clúster (284 pesos).



(c) 3r clúster (100 pesos).



(d) 4t clúster (362 pesos).



(e) 5è clúster (260 pesos).

Figura 5.19: Patrons de pesos (ajustats automàticament) trobats mitjançant l'algorisme EM. El patró de la mediana s'indica amb una línia de punts i la mida de pesos del patró s'indica entre parèntesi. La nomenclatura seguida és la següent: w_1 =PIT.T, w_2 =ENE.T, w_3 =DURL.T, w_4 =DURR.T, w_5 =PIT.C, w_6 =ENE.C, w_7 =MFC.C, w_8 =PosInEG.L, w_9 =PosInSyl.L, w_{10} =PosInWord.L, w_{11} =Prev.L, w_{12} =Next.L, w_{13} =POS.L, w_{14} =Stress.L.

Patrons trobats

A la figura 5.19 es mostren els patrons de pesos obtinguts després d'aplicar l'algorisme EM per cercar 5 patrons de pesos diferenciats (5 és el nombre de grups fixat per les mètriques no conservadores). En termes generals, es pot apreciar que en els patrons (grups), els vectors de pesos ponderen un nombre petit de subcostos (entre 1 i 4) deixant la resta pesos a zero (p.ex, a la figura 5.19(e) només PosInEG.L, PosInWord.L, PosInSyl.L i POS.L adopten valors destacats, esdevenint la resta de pesos insignificants). Els patrons, en línies generals presenten un comportament prou diferent entre sí amb excepció de 5.19(a) i 5.19(d), que són força semblants. El nombre d'instàncies per cada clúster també resulta força equilibrat (entre 100 i 362). La gran majoria de patrons destaquen valors alts en els pesos dels subcostos acústics amb excepció del 5.19(e) que dóna una alta importància a la posició en la paraula i la posició en el grup d'entonació. Aquest comportament resulta normal degut a que els subcostos lingüístics adopten un espai de cerca fortament discret $\{0, 0.5, 1\}$, per tant no s'assoleix la normalitat i la continuïtat que requereix la regressió lineal per modelar les distàncies cepstrals. Aquest fet suposa un nou repte per l'aiGA a nivell teòric, degut a que l'adequació del marc d'ajust interactiu i evolutiu proporciona, en teoria, una eina eficaç per superar les limitacions del modelat lineal.

5.5.4 Ajust de l'algorisme de classificació

Un cop s'han trobat els grups en els que es poden dividir els pesos obtinguts mitjançant ajust automàtic, s'han d'associar aquests patrons de pesos al conjunt de predictors (característiques lingüístiques i fonètiques) per a poder associar cada patró de pesos a cada context/especificitat lingüística i fonètica. Tal i com s'ha esmentat en diversos punts d'aquesta tesi doctoral - apartats 2.1.3, 2.2.2 i 2.4.2 -, el problema de la classificació dins la síntesi de la parla s'ha tractat majoritàriament mitjançant arbres de decisió (Black i Taylor, 1997*a*; Donovan, 2000; Ostendorf i Bulyko, 2002; Navas *et al.*, 2002*a*; Yamagishi *et al.*, 2004). La tria d'aquest d'algorisme ve donada pel seu bon funcionament en reconeixement automàtica de la parla (Ostendorf i Bulyko, 2002). En l'apartat 5.5.1 s'ha esmentat que els dos principals problemes que plantegen els arbres de decisió són el *i*) el *clustering* predictiu i *ii*) l'equilibrat excessiu. Ambdós problemes restringeixen la cerca de clústers d'una manera poc natural. Amb la cerca de grups mitjançant EM s'ha aconseguit superar el problema del *clustering* predictiu tanmateix encara resta el problema de l'equilibri en l'arbre de decisió (Kominek i Black, 2005).

L'algorisme *wagon* (Black i Taylor, 1997*a*) permet configurar el nivell d'equilibri (mi-

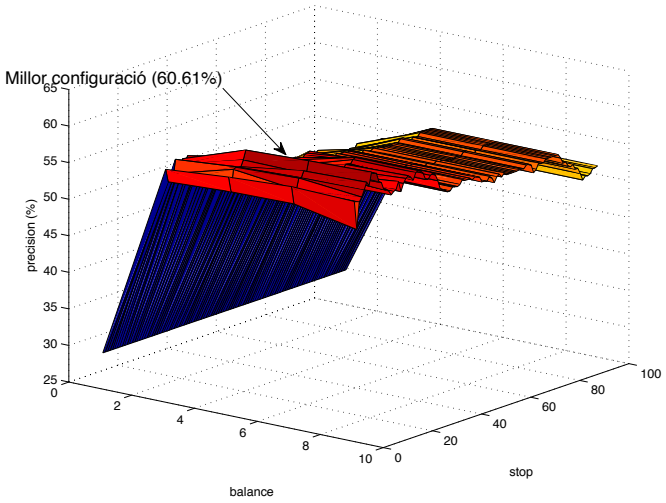


Figura 5.20: Escombrat de valors (*stop-value* i *balance*) per determinar el *balance factor* i *stop value* del *wagon*(Black i Taylor, 1997*a*) emprat. La millor posició (60.61%) s'ubica per *stop value*=8 i *balance*=7.

	1	2	3	4	5	Corr.	Total	% Encert (Prec.)
Clúster 1	**96**	23	17	29	45	96	210	45.71 %
Clúster 2	17	**196**	10	56	5	196	284	69.01 %
Clúster 3	25	18	**39**	14	4	39	100	39.00 %
Clúster 4	22	62	12	**265**	1	265	362	73.20 %
Clúster 5	4	0	1	1	**254**	254	260	97.69 %
Corr.	96	196	39	265	254	850		
Total	164	299	79	365	309		1216	
% Cob. (Recall)	58.54%	65.55%	49.37%	72.60%	82.20%			***69.90%***

(a) Matriu de confusió.

Núm. Clúster	Precision	Recall	*F-measure*
Clúster 1	0.457	0.585	0.513
Clúster 2	0.690	0.656	0.672
Clúster 3	0.390	0.494	0.436
Clúster 4	0.732	0.726	0.729
Clúster 5	0.977	0.822	0.893
Mitjana	0.649	0.657	0.649

(b) Coeficients *F-measure* obtinguts al tractar el problema de classificació multi-classe com un problema de classificació binària.

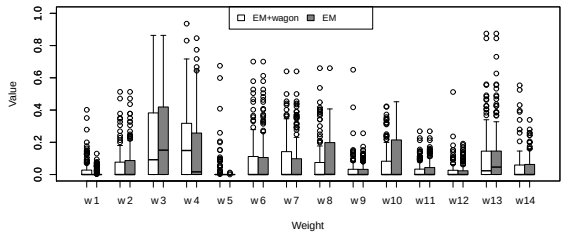
Figura 5.21: Estadístics obtinguts en aplicar l'algorisme *wagon*, prèviament ajustat (*stop value*=8 i *balance*=7), per associar les característiques de predicció als patrons de pesos trobats per l'algorisme EM.

da semblant d'instàncies en les fulles) que s'infereix en l'arbre de decisió en funció d'un percentatge del nombre total de dades a dividir. El valor d'aquest percentatge ve donat per dos paràmetres determinats que són *stop value* i *balance factor* (King *et al.*, 2002). Els paràmetres funcionen de la manera següent: a l'hora de determinar si una node concret s'ha de seguir dividint es parteix el nombre d'instàncies d'entrada pel *factor de balanceig*. Si el valor obtingut és més gran que el nombre d'instàncies màxim (*stop value*) per fulla s'empra aquest valor com a llindar, evitant així la subdivisió d'aquells nodes que tenen prou consistència per sí sols etiquetant-los com a fulles (King *et al.*, 2002).

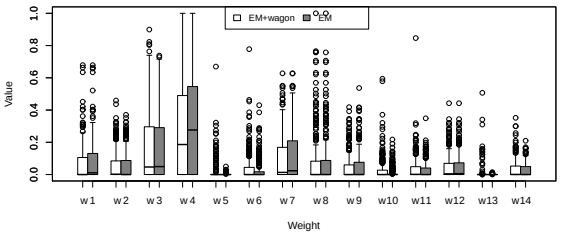
Per a determinar els valors de *stop value* i *balance factor* del problema es realitza un escombrat (figura 5.20) exhaustiu de les diferents possibilitats (acotades dins un interval) mitjançant un procediment de 10-*fold cross-validation*. Al final es determina que la configuració que realitza una millor associació (millor percentatge d'encerts) de la informació contextual (lingüística i fonètica) amb els pesos és $balance = 7$ i $stop = 8$, garantint fulles d'almenys 8 instàncies o una setena part ($1/7$) dels vectors de pesos disponibles en l'entrenament.

L'arbre obtingut es mostra en les figures 5.23 i 5.24. Aquesta metodologia de classificació ofereix un encert de l'ordre del 70% segons la matriu de confusió mostrada en la taula de la figura 5.21(a). Si es vol analitzar el percentatge d'encert amb els paràmetre típics d'un problema de classificació binària s'obté un índex d'exactitud *F-measure* (Hripcsak i Rothschild, 2005) de 0.649 (*macro-averaged*). Els resultats es consideren acceptables pel problema plantejat ja que no es vol trobar una solució exacta al problema dels pesos sinó quins contextos lingüístics i fonètics s'associen a diferents patrons de pesos a l'hora de realitzar la selecció d'unitats.

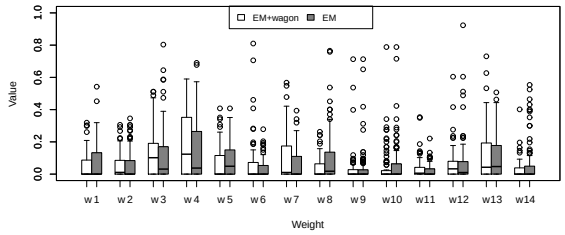
A la figura 5.22 es mostren els canvis que sofreixen els patrons originals un cop modelades mitjançant les característiques lingüístiques i fonètiques dels pesos. Els canvis dels patrons després de l'associació (CART) no resulten molt exagerats respecte els patrons de l'EM, amb excepció del primer patró (veure figura 5.22(a)), on w_4 (DURR.T) passa de tenir un valor de mediana proper a zero a ser el pes amb més importància dins del clúster 1. Això succeeix perquè el *wagon* ha estat incapaç de trobar la combinació de característiques fonètiques i lingüístiques que permetien identificar aquest patró de pesos de manera robusta. Aquesta mancança comporta un canvi en la mida dels clústers, sobretot en els clústers 1 i 3, on existeix una forta confusió entre ells (veure la matriu de confusió de la figura 5.21(a)). Com a efecte col·lateral d'aquest canvi es pot observar com el clúster 5 augmenta de mida considerablement en incorporar 45 combinacions de pesos del clúster 1, no obstant això el seu patró de pesos no se'n veu afectat.



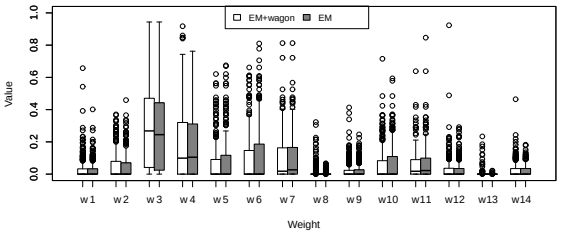
(a) 1r clúster (210→164 pesos).



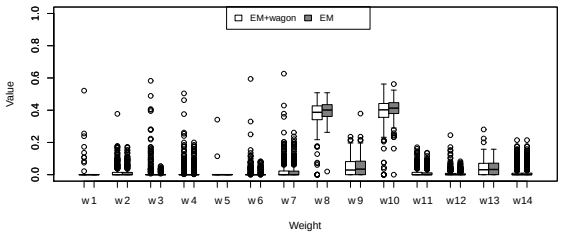
(b) 2n clúster (284→299 pesos).



(c) 3r clúster (100→79 pesos).



(d) 4t clúster (362→365 pesos).



(e) 5è clúster (260→309 pesos).

Figura 5.22: Comparació dels patrons de pesos trobats amb l'arbre de decisió generat a partir de l'algorisme EM (EM+wagon) amb els patrons EM originals. El patró de la mediana s'indica amb una línia de punts i la mida de pesos del patró s'indica entre parèntesi. La nomenclatura seguida és la següent w_1 =PIT.T, w_2 =ENE.T, w_3 =DURL.T, w_4 =DURR.T, w_5 =PIT.C, w_6 =ENE.C, w_7 =MFC.C, w_8 =PosInEG.L, w_9 =PosInSyl.L, w_{10} =PosInWord.L, w_{11} =Prev.L, w_{12} =Next.L, w_{13} =POS.L, w_{14} =Stress.L.

5.6 Consens entre diferents models d'usuari aiGA a partir de models latents

5.6.1 Limitacions del model de consens actual

El model plantejat en el capítol 4 per obtenir els pesos consensuats a partir dels models aiGA planteja, l'inconvenient de la manca d'un model robust que tingui en compte totes les avaluacions dels usuaris per obtenir una solució de consens. En aquest sentit, el model hauria de considerar com evoluciona el *fitness* obtingut de l'ordre complet en les diferents variacions de pesos i no simplement fer un amitjanat dels pesos millors (dominància $>$ 90%) de cada usuari.

Per afrontar aquesta qüestió, aquest apartat proposa l'aplicació de models latents per tal d'obtenir una representació de l'espai de cerca multiusuari a nivell de genotip. No obstant, no es pot estudiar amb detall l'impacte dels models latents fins que els usuaris realitzin les diferents evolucions interactives. En aquest apartat només s'exposa la discussió teòrica del problema deixant l'estudi real de la seva millora per l'apartat d'experimentació pròpiament dit (apartat 5.8.1).

La primera aproximació al problema de consensuar els criteris dels diferents usuaris (apartat 4.6.2) estableix la metodologia de consens següent: per cada frase d'ajust, un cop l'han optimitzat els diferents usuaris, es seleccionen un conjunt de vectors de pesos candidats a ser definitius, que són els pesos que presenten una major dominància, d'almenys $\geq$ 90% (ordre complet induït de l'apartat 3.4.3). Convé recordar que aquesta taula es calcula a partir de les dominàncies dels grafs dirigits que representen les preferències dels diferents usuaris (ordre parcial). Un cop seleccionats, els pesos candidats conformen una matriu que permet visualitzar la distribució de valors per cada pes en els millors genotips i així seleccionar la seva mediana per obtenir la combinació de pesos final.

A la figura 5.25 es mostra un exemple gràfic del procés de consens dels pesos finals per frase un cop seleccionats els pesos de major dominància. Com es pot observar a la figura, el consens dels pesos es pot entendre com una metodologia 1-NN (*Nearest Neighbour*) on el veïnatge el marca el centroide (calculat a partir de les medianes de cada pes).

Tanmateix, aquesta aproximació presenta certes mancances: en aplicar el criteri de mediana es perd informació rellevant provinent de l'evolució, com ara la correlació dels pesos entre d'ells o la multimodalitat de l'espai de cerca. És a dir, no permet determinar si els pesos estan correlats entre ells o bé si un usuari ha convergit a múltiples solucions. Aquesta informació, malgrat estar reflectida en les mètriques ρ i τ no s'utilitza de cap manera en el

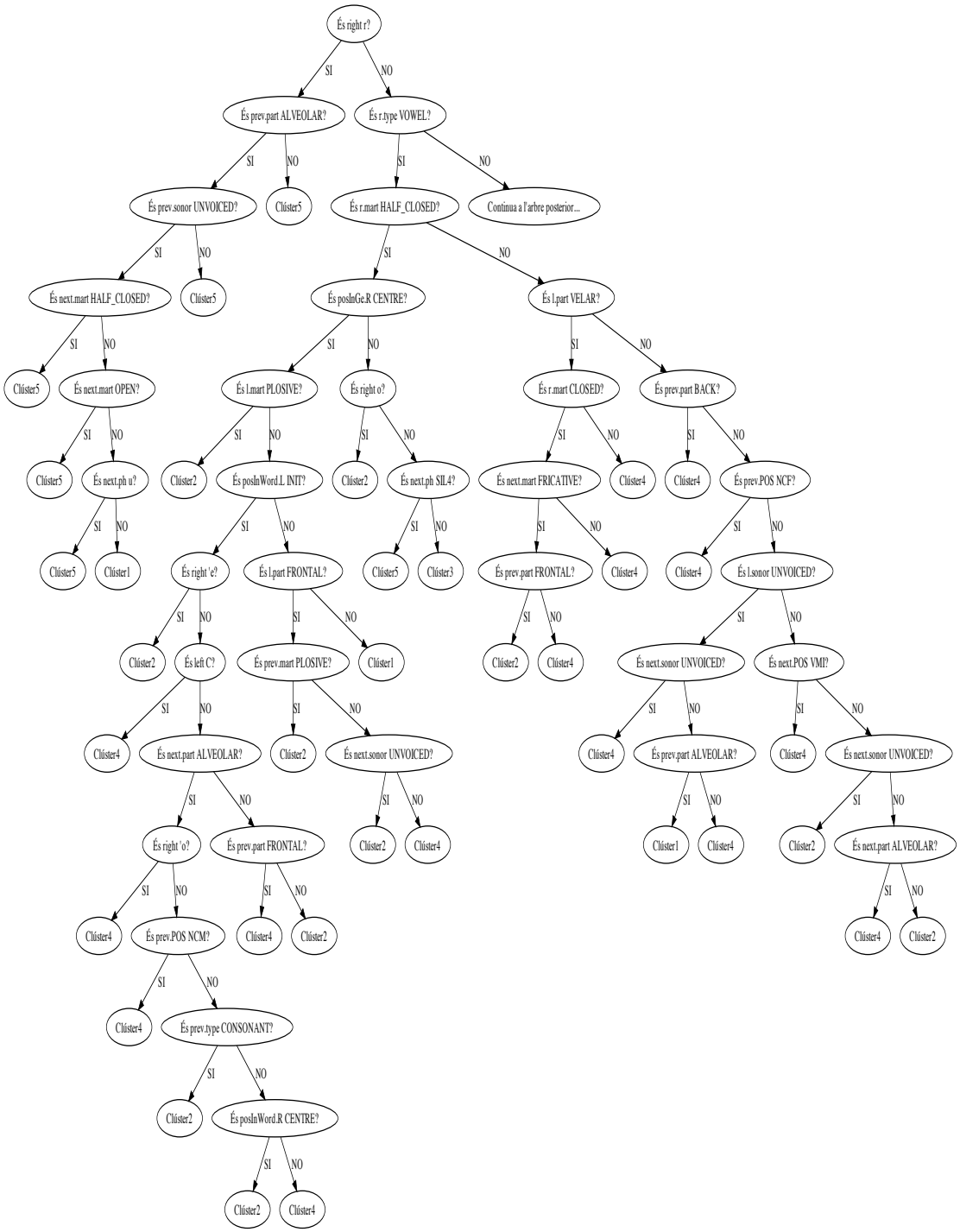


Figura 5.23: Part esquerra de l'arbre de classificació emprat per associar el context lingüístic i fonètic a un clúster de pesos en el corpus *url_dav_es*.

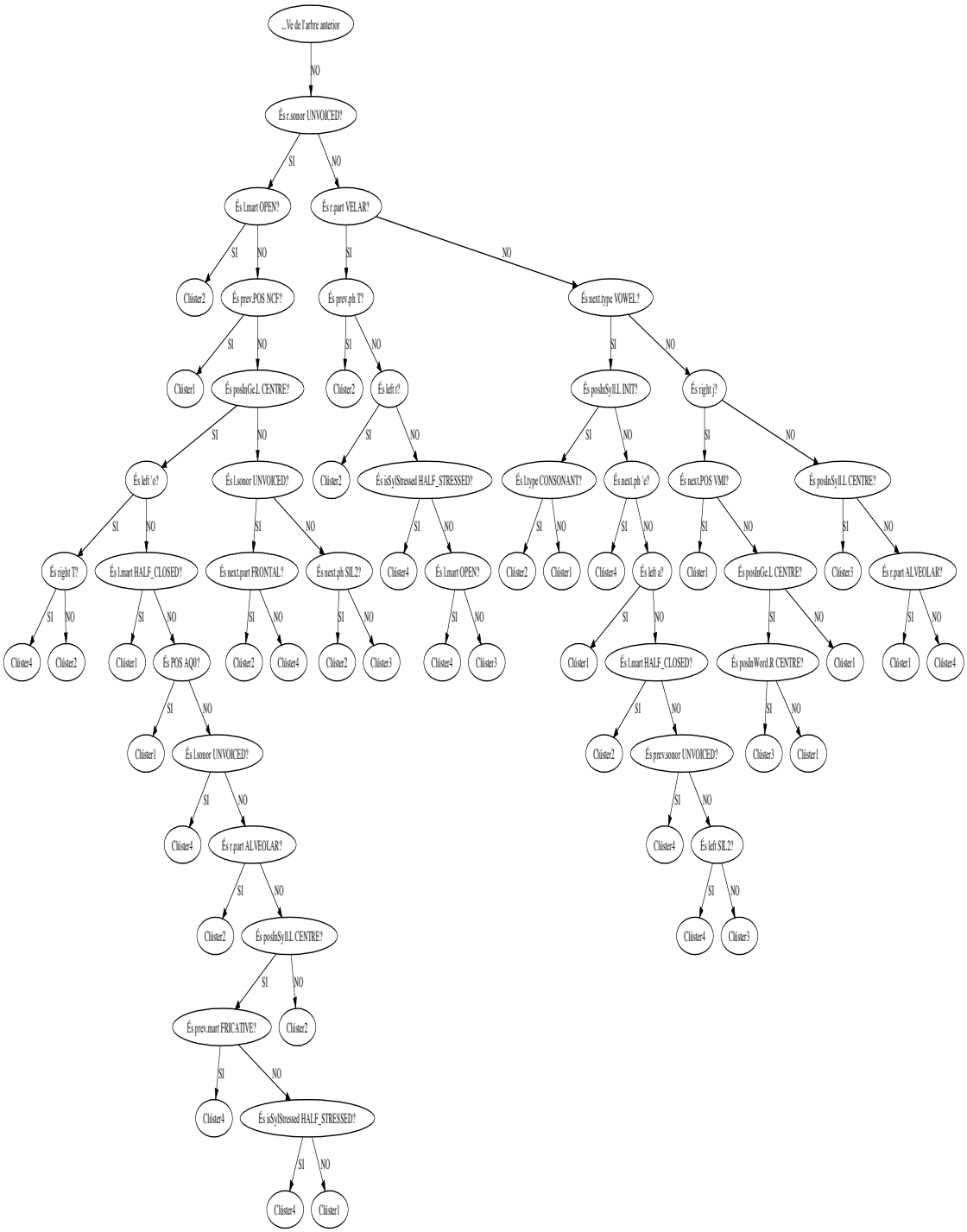


Figura 5.24: Part dreta de l'arbre de classificació emprat per associar el context lingüístic i fonètic a un clúster pesos en el corpus *url_dav_es*.

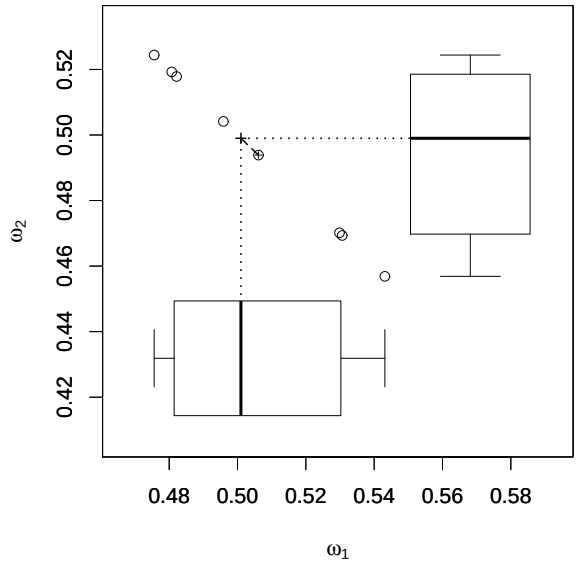


Figura 5.25: Exemple del consens de pesos entre els diferents usuaris mitjançant medianes. A efectes de fer l'exemple més entenedor es simplifica la dimensionalitat a només dos pesos: w_1 i w_2. Els valors de pesos w_1 i w_2 s'ubiquen en una línia recta degut a la restricció $\sum_i w_i = 1$.

consens dels pesos finals.

És en aquest punt on es fa necessari que el consens de pesos es realitzi sobre un model prèviament entrenat que englobi les relacions dels pesos entre ells i el *fitness*. En altres paraules, el problema de l'aproximació mitjançant mediana és la manca d'un model general que consideri els diferents paisatges (*landscapes*) de *fitness* segons els pesos avaluats per cada usuari, i així poder-ne treure una optimització consensuada.

La idea de consensuar criteris de diferents usuaris ja ha estat estudiada pels creadors de l'aiGA (Llorà *et al.*, 2008). La seva aproximació ataca el problema a partir de la construcció d'un graf general que agrupi els grafs parcials que han constituït els diferents usuaris per així aprofitar la representació de les avaluacions pròpia l'aiGA. Aquest graf general l' anomenen conjunt d'avaluacions paral·leles (*parallel evaluation ensemble*). El seu plantejament és que el model global de consens corregeixi les contradiccions dels usuaris, siguin amb ells mateixos o amb els altres usuaris. Tanmateix, aquestes contradiccions continuen quedant representades en forma de cicles en el graf general que s'han de trencar mitjançant les heurístiques guiades exposades en l'apartat 3.4.7 (Llorà *et al.*, 2005*a*). Cal esmentar que el seu (Llorà *et al.*, 2008) model de consens es basa en fenotips (en el nostre cas en les síntesis d'àudio), fet que impedeix cercar relacions intrínseques en l'estructura dels genotips (p.ex.

correlacions lineals o d'ordre superior) que puguin donar una idea més acurada de quin és l'espai de cerca real i multiusuari a nivell de vectors de pesos.

Com alternativa, en aquesta tesi, es proposa cercar un model general basat en genotips que permeti estudiar les relacions intrínsiques dels pesos entre ells mateixos i alhora que permeti modelar l'espai de cerca en funció dels vectors de pesos (genotips) i no dels fenotips. Per tenir aquest enfocament es vol estudiar la idoneïtat d'emprar models latents (Gibson, 1959) per obtenir un model perceptiu que englobi els criteris de diferents usuaris a nivell de genotip.

Els models latents, emprats àmpliament en l'àmbit de la psicologia i les ciències socials (Bollen, 2002), capturen les relacions intrínsiques d'un conjunt de dades transformant (o projectant) l'espai original a un nou espai de dimensionalitat reduïda (1 o 2 dimensions) que ajuda a comprendre l'estructura original. Cal remarcar el fet que els models latents permeten tractar dades complexes, incertes i també amb soroll (Fornells, 2007). Aquest espai nou, així com les variables que el conformen, es coneix generalment com espai latent o ocult (Bishop *et al.*, 1998).

L'exemple més conegut d'aquesta aproximació és l'anàlisi factorial (tècnica semblant a l'anàlisi per components principals o PCA), el qual realitza l'esmentada transformació mitjançant una combinació lineal de les dimensions de l'espai original. Tanmateix els models latents basats en transformació lineal són restrictius en el sentit que només poden trobar comportaments/relacions lineals en les dades, obviant dependències d'ordre superior. Es fan necessaris, doncs, models que no presentin aquesta restricció lineal. En aquest sentit destaquen els models latents adaptatius com a mètodes de modelat latent segons aproximació no lineals (Kohonen, 1982).

Aquests models latents adaptatius poden, de manera automàtica, generar un espai latent (mapa) uni o bidimensional que reflecteixi les propietats presents en el conjunt de dades de l'espai inicial. La propietat més important és que si en l'espai original les dades estan mètricament relacionades d'una manera estructurada, la mateixa estructura es reflectirà en relacions espaials dins l'espai latent mitjançant les relacions de veïnatge (Kohonen, 1982). A més, aquesta propietat es manté independentment de l'ordre inicial d'aquestes dades (organització de les dades dins l'espai). Per dir-ho d'una manera més comprensible (Gibson, 1959), els models latents adaptatius són capaços de formar imatges de relacions entre les variables.

Cal afegir que les relacions no lineals entre les dades del conjunt d'entrada han estat un problema difícil de resoldre en termes matemàtics. Normalment aquests mapes o models s'usen en la seva variant bidimensional, adoptant diferents dimensions o topologies

(distribució hexagonal o rectangulars de les variables latents en l'espai bidimensional) en funció de les dades d'entrada. A la figura 5.26 es mostra un exemple de modelat latent. Concretament, a la figura 5.26(a) es mostra com quedarien distribuïts els patrons de pesos de tots els usuaris d'una prova concreta, ja agrupats, en un espai latent rectangular de 4x4. A la figura 5.26(b) es poden veure els valors de *fitness* associats a cada vector de pesos seguint exactament la mateixa distribució.

Aquests models tenen diferents aplicacions dins del propi àmbit de l'aprenentatge artificial, entre les que destaquen principalment el *clustering* de trets rellevants en les dades (veure apartats 4.4 i 5.5) o la representació gràfica del coneixment d'una manera comprensible (Gibson, 1959). Altres autors han emprat la capacitat de *clustering* dels models latents per millorar l'eficiència de l'aprenentatge automàtic (concretament raonament basat en casos – (Fornells, 2007)).

Juntament amb aquestes utilitats també destaca l'imputació de dades no conegudes o perdudes (*missing data imputation*) (Fessant i Midenet, 2002; Vellido, 2006). A la figura 5.26(c) es poden veure els *fitness* associats a cada patró de pesos prèviament agrupat (figura 5.26(a)).

La imputació de dades no conegudes permet realitzar un pas més en l'aplicació de models latents a l'hora de trobar comportaments de les diferents variables entre si o dels diferents usuaris. Aquest pas ve donat per la capacitat dels mètodes latents adaptatius de ser emprats com a models predictius (Fessant i Midenet, 2002) o regressors per tal de corregir dades absents o errònies i així inferir un ordre més natural en les dades. Aquesta característica permet tractar, d'una manera no heurística, el problema de les contradiccions entre diferents usuaris o d'un usuari amb sí mateix: a l'obtenir un model global es poden corregir dades incorrectes, definides com aquelles que presenten més variació entre la predicció del model i el seu valor real. Alhora, el mateix model pot servir per identificar aquells patrons de pesos que són bons dins del model consensuat per així obtenir-ne la combinació de pesos guanyadora o final considerant el centroide del grup que presenti els millors *fitness*. A la figura 5.26(d) es mostra un exemple de predicció del *fitness* mitjançant l'imputació de dades no conegudes.

Algorisme 5.7 Obtenció dels pesos finals mitjançant un model latent multiusuari.

procedure $resultExtraction(\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_\mathcal{U})$

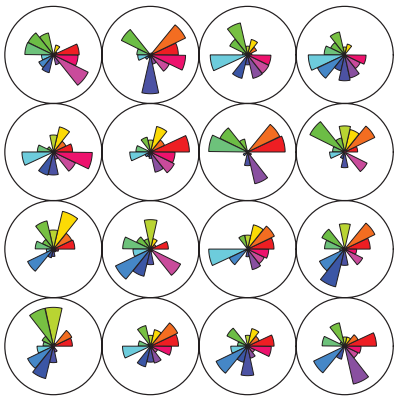
1: Determinar la topologia (t) i dimensions ($m \times n$) del model latent mitjançant la minimització de l'error RMSE en la predicció (*missing data imputation*) de valors de *fitness*. La minimització es realitza mitjançant l'escombrat de diferents topologies i dimensions segons l'estratificació *10-fold cross validation*.
2: Entrenar un model latent $\mathcal{M}(t, m, n)$ a partir de les taules d'ordre complet dels diferents grafs $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N$
3: Determinar mitjançant *clustering* els patrons $\mathcal{P} = (p_1(\mathcal{M}), p_2(\mathcal{M}), \ldots, p_P(\mathcal{M}))$ rellevants en l'espai latent del model $\mathcal{M}$
4: Seleccionar el patró p_i amb el valors de *fitness* més alt: $p_{win} = max_{i=1}^{P} fitness(p_i)$
5: Seleccionar la combinació de pesos $w_{win} \in p_{win}$ que presenti la distància més propera al centroide(p_{win}) dins l'espai latent
6: retornar w_{win}

La combinació de les diferents capacitats dels models latents: *i*) *clustering* de patrons multiusuari i *ii*) la correcció de valors de *fitness* mitjançant regressió, permeten obtenir els pesos finals segons la metodologia que es detalla a l'algorisme 5.7.

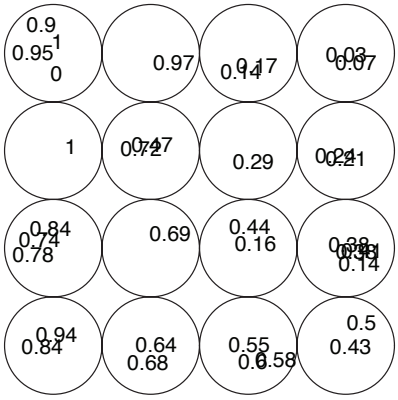
Els dos sistemes més coneguts del modelat latent adaptatiu són els mapes autoorganitzatius (*Self-Organizing Maps* - SOM o mapes de Kohonen) (Kohonen, 1990) i els mapes topogràfics generatius (*Generative Topographic Mapping* - GTM) (Bishop *et al.*, 1998), els quals es detallen a continuació.

5.6.2 Mapes autoorganitzatius

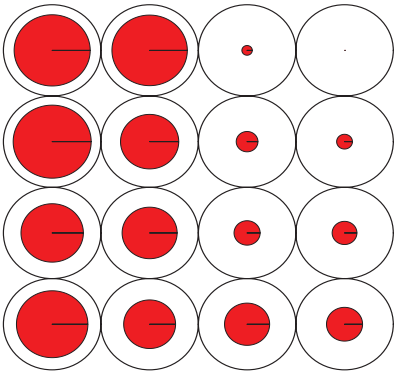
Els mapes autoorganitzatius (SOM) (Kohonen, 1990) són un model de xarxes neuronals de dues capes, una capa en l'espai original de les dades (multidimensional) i l'altre en l'espai latent (uni o bidimensional). La capa d'entrada té tants nodes o neurones com dimensions té l'espai d'entrada (en aquest cas, 14 pesos) i s'encarrega de transformar i projectar l'informació d'entrada a la capa de la xarxa que es troba en l'espai latent. D'altra banda, la capa de l'espai latent s'encarrega de tractar la informació que li transmeten les neurones de la capa d'entrada i definir agrupacions de les dades en l'espai latent. Les neurones d'entrada no estan connectades entre elles. En canvi, les neurones d'entrada estan connectades amb totes les neurones de sortida mitjançant un valor f (ponderació) determinat. La combinació dels diferents valors f en una neurona de sortida s'anomena vector director o vector de



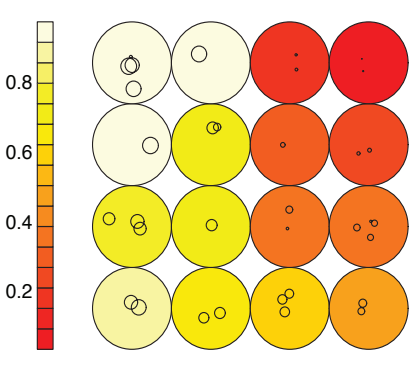
(a) Distribució dels diferents patrons de pesos dels diferents usuaris, un cop agrupats, en l'espai latent un cop obtingut el model.



(b) Distribució dels valors de *fitness* en l'espai latent segons el mateix model latent.



(c) *Fitness* associat a cada patró segons el modelat latent.



(d) *Fitness* predit per cada combinació de pesos segons el model latent mitjançant la tècnica de *missing data imputation*. La mida dels cercles es proporcional al *fitness* predit. El color del cercle determina el *fitness* del patró.

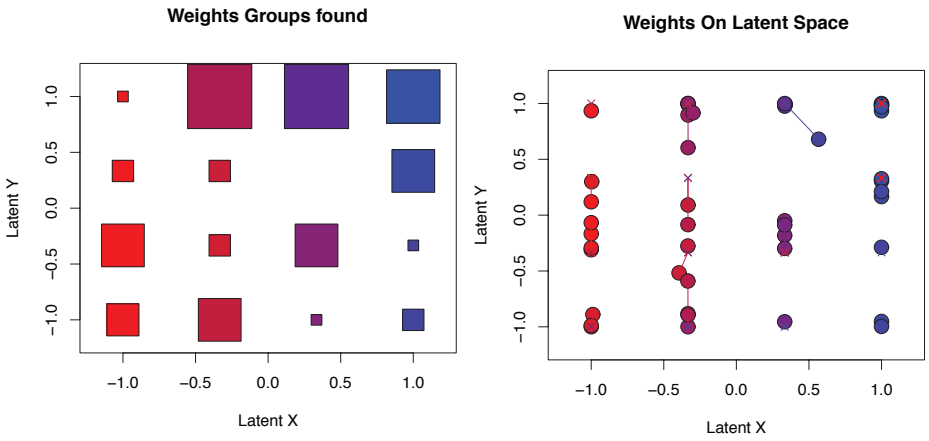
Figura 5.26: Punts de vista diferents del mateix modelat latent dels pesos de 3 usuaris per una mateixa prova aiGA (frase "*poco a poco*" – clúster 2 – taula 5.14) del corpus *uvig_dav_es*.

referència i permet ponderar quines dades d'entrada activen una neurona en l'espai latent. Les neurones de la capa de sortida, estan connectades entre elles mitjançant unes connexions de veïnatge, que determinen la zona d'influència de cada neurona. Val a dir que cadascuna de les neurones està associada únicament i exclusiva amb les seves veïnes dins l'espai latent (8 connexions segons topologia rectangular i 6 connexions segons topologia hexagonal). La topologia de la xarxa (recangular o hexagonal) es tria en funció de la precisió de les dades a mapar. La figura 5.26, mostra un exemple de mapes autoorganitzatius. La implementació de SOM que s'ha emprat en aquesta tesi doctoral és la *XY-Fused network* (XYF) (Melssen *et al.*, 2006) degut a la seva adaptació per resoldre problemes de regressió (en l'aiGA es vol realitzar una regressió entre els diferents *fitness* i els pesos obtinguts).

5.6.3 Mapes topogràfics generatius

Una de les motivacions de la definició dels mapes topogràfics generatius (GTM) (Bishop *et al.*, 1998) va ser la de proporcionar una alternativa als mapes autoorganitzatius (Kohonen, 1990). Malgrat que l'aproximació SOM ha tingut molta acceptació en aplicacions pràctiques (p.ex. tecnologies audiovisuals, reconeixement/processament de la parla, medicina, etc. (Oja *et al.*, 2003)), conté certes mancances observades pel mateix Kohonen (1995). Aquestes mancances venen donades pel seu comportament no estocàstic. Concretament les seves limitacions són (Bishop *et al.*, 1998): *i*) l'absència d'una funció de cost que penalitzi els vectors de referència en funció de la seva capacitat, *ii*) la manca d'una base teòrica que permeti escollir la progressió del factor d'aprenentatge (interval màxim en el que es pot desplaçar la neurona a mesura que s'entren dades en el model) així com els paràmetres de veïnatge i assegurar un ordre topogràfic en l'espai latent, *iii*) la manca d'un índex de convergència que determini la durada de l'entrenament i *iv*) el fet que el model no defineix una funció de densitat de probabilitat sobre la distribució dels valors dins del clúster. Aquests problemes venen donats per l'origen heurístic i estocàstic de l'algorisme SOM (Bishop *et al.*, 1998). A continuació s'explica com els GTM poden superar aquestes limitacions.

GTM es basa en una combinació de gaussianes (*Gaussian Mixture Model* — GMM) restringida (en forma de topologia rectagonal o hexagonal prèvia). En el cas de GTM, els paràmetres d'aprenentatge i el valors del vector de referència s'estimen segons l'algorisme EM (Moon, 1996) (veure apartat 5.5.3). El problema del models latents heurístics (SOM) s'esdevé per la incapacitat d'establir un ordre a priori de les dades per cada neurona (o clúster) mitjançant una estimació de dades. En canvi, en GTM el conjunt de punts latents s'entén com una malla. Cada punt de la malla és una funció de distribució gaussiana cir-



(a) Gaussianes (rectangles) de la malla GTM posicionades segons un ordre establert a priori en l'espai latent. El color de la gaussiana indica el valor de *fitness* del seu centroide (vermell: $fitness \approx 1$ / blau: $fitness \approx 0$). La mida del quadrat determina el nombre d'exemples d'entrenament que són cobertes per la gaussiana.

(b) Posicionament dels diferents vectors de pesos d'entrenament en l'espai latent. Les aspes indiquen el centroide (gaussiana) en l'espai latent i les línies indiquen la pertinença. El color de l'exemple indica el seu valor de *fitness* (vermell: $fitness \approx 1$ / blau: $fitness \approx 0$).

Figura 5.27: Exemple de modelat latent mitjançant GTM (topologia quadràtica amb malla de 4×4). La figura (a) mostra les diferents gaussianes del GTM posicionades en l'espai latent mentre que la figura (b) mostra les variables d'entrada projectades en l'espai latent. El GTM modela els pesos de 3 usuaris per una mateixa prova aiGA (frase "*clasificadas así*" — clúster 4 — taula 5.14) del corpus *uvig_dav_es*.

cular amb la seva correspondència equivalent (projecció) a l'espai multidimensional d'entrada. Aquesta projecció es realitza a través d'unes funcions no lineals de canvi de base ponderades. Llavors, l'ordre a priori específic entre les gaussianes ve determinat per l'estimació de les dades que realitza l'algorisme EM, permetent que la malla es deformi per cobrir les dades d'entrada. Una primera aproximació del GTM per consensuar el criteri dels diferents usuaris en aiGA es va estudiar a Formiga i Alías (2007). A la figura 5.27 es mostra un exemple mitjançant modelat GTM. En aquesta tesi s'ha emprat la implementació GTM de Bishop *et al.* (1998). Aquesta variant no proporciona cap metodologia per realitzar la imputació de dades no conegudes. A tal efecte, s'utilitza un algorisme inspirat en la metodologia proposada per Fessant i Midenet (2002) que permet realitzar regressió en un model latent. El procés es detalla a continuació:

i) Presentació d'un vector incomplet w_i (sense *fitness*) a l'entrada del model GTM $\mathcal{M}$.

ii) Determinació de la posició p d'aquest vector en l'espai latent segons les probabilitats a posteriori (responsabilitats) d'aquest vector w_i respecte cadascuna de les gaussianes del model latent $\mathcal{M}$. S'omet la dimensió del valor no conegut (el *fitness*) de la mitjana μ de cada gaussiana.

iii) Obtenció del *fitness* (valor no conegut) $f(w_i)$ associat a w_i calculat com a la combinació lineal dels diferents centroides μ ponderats segons les probabilitats a posteriori (responsabilitats) de cada gaussiana respecte la posició p.

iv) Retorna $f(w_i)$

En l'apartat d'experiments (apartat 5.8.1), s'avalua la idoneïtat perceptiva d'emprar models latents (GTM/SOM) per consensuar les solucions de diferents usuaris.

5.7 Ajust perceptiu dels pesos mitjançant aiGA

5.7.1 Disseny de les proves

Abans d'entrar amb detall a l'ajust de pesos perceptiu mitjançant aiGA es repassa breument la proposta proposada durant les contribucions d'aquest capítol. La idea es ajustar perceptivament els pesos que ponderen els diferents subcostos dins la funció de cost respectant la especificitat fonètica i contextual de cada unitat en el moment de la selecció. Convé recordar que per realitzar l'ajust en un entorn real de selecció d'unitats s'ha passat

Clúster	Frase	Unitats Variables	Unitats Totals
1	ing***res***ad***os en el hos***pital	8	21
1	p***roc***edim***ien***t***os en él estable***cidos	12	30
1	***trab***ajo ***en el ex***tr***anj***ero	10	22
1	v***ez en onc***e	6	9
2	***des***cubrí***an poco a poco y bajo***s	12	25
2	pu***nto de apoyo y*** de	10	15
2	***poco a poco l***a	6	11
2	***baj***ur***a como de arras***tre	13	19
3	la r***eunión***	5	8
3	Ca***rmen, He***rt, apena***s l***e	5	9
3	c***amión,*** que	5	9
3	alc***alde, gon***zalo	5	15
4	***adjudicado*** es***t***e	11	14
4	co***cina en*** bl***an***co	8	14
4	***adjudicad***a ***a*** la	8	10
4	***c***las***ificad***as, ***así***	9	12
5	se regis***tró***	3	9
5	suspendía siem***pre***	2	8
5	de noviem***bre***	2	10
5	conserva siem***pre***	2	11

Taula 5.14: Frases escollides per realitzar les proves interactives usant aiGA dissenyades pel corpus *uvig_dav_es*. Amb negreta es destaquen les unitats variables (que admeten selecció d'unitats) respecte les unitats portadors (fixes en tot l'ajust).

d'un corpus de 8 minuts (*url_fer_ct*) a un corpus de 1.9 hores (*uvig_dav_es*) i s'han augmentat els subcostos de 7 a 14 (incorporant els subcostos lingüístics). Posterioment s'ha realitzat l'ajust automàtic d'aquests pesos mitjançant distàncies cepstrals i emprant MLR/NNLS en comptes de GA. Cal tenir en compte que s'ha incorporat informació contextual en l'ajust automàtic dels pesos passant de tenir pesos a nivell d'unitat a tenir pesos a nivell de subunitat contextualitzada. L'agrupació dels pesos obtinguts automàticament s'ha separat en dues etapes, una de *clustering* pròpiament dita i una d'associació o mapat que associa els diferents contextos (fonètic i lingüístics) a cada patró de pesos.

Un cop obtinguts els patrons de pesos segons la metodologia d'ajust automàtica s'han d'escollir quines són les frases que millor representen els esmentats patrons (veure figura 5.22). Per decidir les frases fonèticament balancejades que conformen cada grup (seguint els experiments previs s'escullen 4 frases per clúster) s'aplica l'algorisme de màxima entropia explicat en l'apartat 4.4.1. Les frases del corpus són majoritàriament extenses (≈ 80 unitats per frase) provocant que l'usuari hagi de mantenir el criteri de qualitat durant aproximadament 10 segons, impedint una avaluació clara de les frases. En aquest sentit, es modifica lleugerament l'algorisme per cercar també dins els grups d'entonació de les diferents frases i així no introduir fatiga i ambigüitat en l'usuari avaluador que realitzarà l'ajust de pesos.

En el disseny de les proves, es segueix la metodologia dels experiments anteriors (apartats 3.4.8 i 4.6), on es considera la prosòdia original com la idònia per tal de guiar el procés d'ajust pesos i així no veure's afectada pels errors del modelat prosòdic previ (veure 4.2.2). A la vegada, se separen les frases d'ajust en unitats portadores i unitats variables per tal que els usuaris es concentrin en les unitats que són pròpiament del clúster. Les frases escollides es detallen a la taula 5.14, on es pot observar que les unitats usades per l'ajust de cada grup guarden característiques similars entre elles. Aquesta semblança és deguda a la naturalesa fonètica i lingüística del *clustering* previ: per exemple, a la taula 5.14 es pot veure que el clúster 5 concentra les terminacions de grup d'entonació segons el patró oclusiva-líquida-vocal (segons la taula D.1 de l'annex). A nivell estadístic les frases usades per l'ajust de pesos contenen de mitjana 14.05 ± 6.25 unitats, de les quals 7.1 ± 3.52 són variables i canvien en la funció de cost de selecció.

Cal esmentar que l'augment del nombre de variables en l'ajust (de 7 a 14) no afecta la mida de les poblacions a avaluar. Aquest fet ve justificat per l'argumentació de l'apartat 3.4.3 on la mida de la població es defineix segons el nombre de variables a ajustar. En el nostre cas, $\ell + 2 = 14 + 2 = 16$ — on ℓ és la mida mínima de la població per tenir un *fitness* sintètic acceptable, que és la mateixa mida que s'havia emprat en les proves anteriors

(veure apartat 3.4.6).

Seguint la mateix posada a punt que la descrita en l'apartat 3.4.8, les generacions evolutives es composen de 16 combinacions de pesos comparades segons un torneig binari (15 comparacions entre 2 mostres de veu sintètiques a cada iteració). Cada prova evoluciona en un total de 3 generacions ($3 \times 15 = 45$ avaluacions). Val a dir que en tot moment el comportament dels usuaris es monitoritza per mitjà dels indicadors κ, λ, ρ i τ (veure apartats 3.4.7 i 4.5).

5.7.2 Resultats

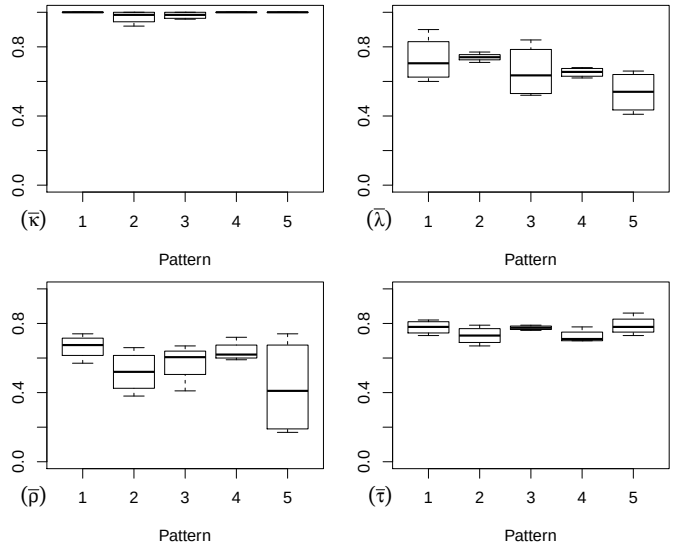
Un total de 8 usuaris experts en síntesi de veu van participar en l'ajust dels pesos corresponent a les unitats del clúster que composen les frases, garantint que cada frase era ajustada almenys per part de 3 usuaris diferents, i a la vegada, que cada usuari ajustava frases de dos clústers diferents. Un cop ajustats els pesos, cada usuari va emprar una mitjana 8.7 ± 3.18 minuts en l'ajust de cada frase. El comportament dels usuaris en el procés d'ajust evolutiu i els patrons de pesos obtinguts es detallen a continuació.

Consistència, ambigüitat, convergència i correlació dels pesos

A la figura 5.28 es detalla l'evolució dels diferents indicadors κ, λ, ρ i τ per clúster 5.28(a) i per frase 5.28(b).

En termes de consistència, totes les frases acaben amb un nivell de consistència alt ($\kappa > 0.97$). Tanmateix, quatre frases (dues del segon clúster i dues del tercer clúster) presenten un lleuger decreixement de la consistència al final del procés evolutiu. Com a conseqüència, es descarta un usuari de la frase 6 ($\kappa = 0.76$), un usuari de la frase 8 ($\kappa = 0.9$), un usuari de la frase 10 ($\kappa = 0.9$ - el mateix que el de la frase 9) i un usuari de la frase 11 ($\kappa = 0.9$). En total, es descarten un 6.66% del total d'ajustos evolutius realitzats.

En termes d'ambigüitat, es pot observar com les frases del clúster 1 i 2 són les que presenten un major índex de certesa en termes generals, encara que els clústers 1, 3 i 5 tenen molta variabilitat en l'índex de certesa depenent de la frase ajustada. Les frases del clúster 5 són les que han resultat més confuses, i per tant ambigües, en el global de les diferents evolucions interactives, destacant el fet que dues d'elles (17 i 20) obtenen un índex de certesa significativament baix ($\lambda < 0.5$). Respecte a la durada efectiva de les proves (l'últim punt d'aportació d'avaluacions - veure apartat 4.6.3), l'última iteració informativa s'ubica aproximadament al voltant de l'avaluació 40.7 ± 3.02.



(a) Indicadors obtinguts detallats per nivell (amitjanats per usuari).

Test id	Clúster	$\bar{\kappa}$	$\bar{\lambda}$	$\bar{\rho}$	$\bar{\tau}$
1	1	1	0.6	0.74	0.82
2	1	1	0.9	0.57	0.73
3	1	1	0.65	0.69	0.76
4	1	1	0.76	0.66	0.8
5	2	1	0.77	0.38	0.75
6	2	0.92	0.71	0.57	0.71
7	2	1	0.74	0.66	0.79
8	2	0.97	0.74	0.47	0.67
9	3	1	0.54	0.41	0.79
10	3	0.97	0.84	0.61	0.78
11	3	0.96	0.52	0.60	0.76
12	3	1	0.73	0.67	0.77
13	4	1	0.62	0.63	0.78
14	4	1	0.67	0.61	0.72
15	4	1	0.64	0.59	0.7
16	4	1	0.68	0.72	0.7
17	5	1	0.46	0.17	0.77
18	5	1	0.66	0.61	0.79
19	5	1	0.62	0.74	0.73
20	5	1	0.41	0.21	0.86

(b) Indicadors obtinguts detallats per frase.

Figura 5.28: Indicadors obtinguts detallats per nivell (a) i per frase (b). On $\bar{\kappa}$ mesura la consistència, $\bar{\lambda}$ mesura la certesa, $\bar{\rho}$ mesura la convergència intra-usuari i $\bar{\tau}$ mesura la correlació inter-usuari.

L'índex de convergència intra-usuari reflecteix la informació d'ambigüitat obtinguda en l'indicador de certesa. En primera instància es pot observar com el baix índex de certesa obtingut en les frases 17 i 20 implica un baix índex de convergència intra-usuari. De fet, dels 6 usuaris (3 per frase) que han realitzat l'evolució de les dues frases, només dos (un per frase) han aconseguit acabar la frase amb un índex de convergència intra-usuari acceptable ($\rho = 0.51$ i $\rho = 0.61$): la resta d'usuaris no han aconseguit dominàncies superiors al 90% impedint obtenir solucions del seu graf. Des d'un punt de vista de clústers es confirma que el cinquè clúster, és el que conté pitjors índexs de convergència ratificant així la informació que aporten els índexs de certesa. Alhora, es confirma que el segon i tercer clústers presenten molta variabilitat en l'índex de convergència, arribant a assolir valors de $\rho < 0.5$. Si a aquest fet se l'hi suma que els dos nivells tenen un índex de certesa (λ) prou acceptable es pot deduir que combinacions de pesos distintes entre elles, són igualment bones per l'usuari de manera robusta esdevenint múltiples solucions per una sola evolució.

Respecte l'índex de correlació dels pesos entre els múltiples usuaris no s'aprecien diferències destacables, fet que constata que els pesos vàlids obtinguts entre els diferents usuaris són prou semblants entre ells.

En global es constata que el patró de pesos que s'obté del clúster 5 és el més difícil per a obtenir pesos de manera perceptiva. Aquest fet està relacionat amb que les frases de l'esmentat clúster són les que menys unitats variables tenen dins el global de la síntesi ($\approx$ 2.25 unitats variables per frase – veure taula 5.14)

Tanmateix la resta de patrons presenten comportaments acceptables en la majoria d'indicadors (> 0.5) malgrat que el nivell 2 i nivell 3 no ofereixin una única solució al problema de l'ajust. A continuació es detallen els pesos obtinguts.

Pesos obtinguts

A la figura 5.29 es poden observar els pesos obtinguts segons MLR/NNLS i l'aiGA. Els pesos de l'aiGA es desglossen segons les diferents tècniques de consens de resultats: Mediana (també anomenada mediana o 1-NN – veure l'apartat 4.6.2) i les dues tècniques basades en models latents (SOM i GTM) explicades en l'apartat 5.6.1. Les correlacions (correlació de cosinus – veure l'apartat 4.4.4) dels patrons de pesos entre les diferents metodologies de consens es poden observar a la taula 5.15 detallades per clúster.

Es pot observar que les diferents metodologies de consens dels pesos de l'aiGA (Median, SOM i GTM) tenen una correlació alta ($corr > 0.88$) entre elles per a tots els clústers.

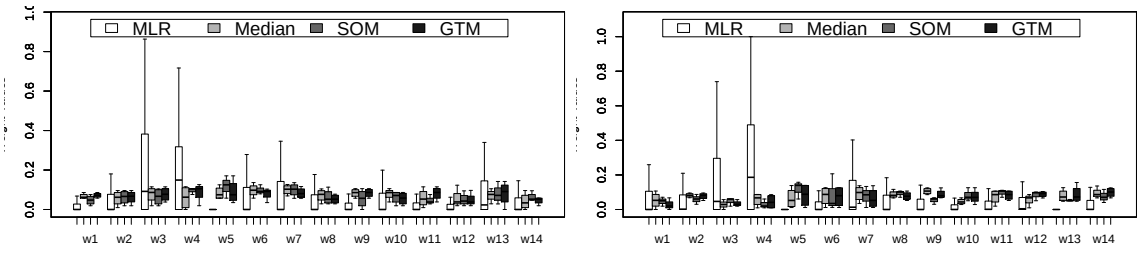
En canvi els pesos automàtics obtenen coeficients de correlació moderats ($corr < 0.55$) respecte l'aiGA. La diferència més notòria es troba en els pesos lingüístics ($w_8 - w_{14}$), on els pesos perceptius prenen valors superiors respecte els pesos automàtics. Aquest fet demostra la capacitat de l'aiGA per tractar amb subcostos fortament discretitzats $\{0, 0.5, 1\}$ a diferència de la metodologia d'ajust de pesos basada en regressió lineal, responent així la qüestió plantejada en l'apartat 5.5.3 sobre la capacitat de l'aiGA d'ajustar subcostos fortament discretitzats.

Mètode/Clúster	*1*	*2*	*3*	*4*	*5*
MLR-Median	0.38	0.29	**0.54**	0.43	0.31
MLR-SOM	0.47	0.14	0.46	0.46	0.33
MLR-GTM	0.49	0.19	**0.51**	0.39	0.32
Median-SOM	**0.95**	**0.88**	**0.95**	**0.94**	**0.94**
Median-GTM	**0.96**	**0.95**	**0.96**	**0.97**	**0.91**
SOM-GTM	**0.95**	**0.93**	**0.97**	**0.97**	**0.94**

Taula 5.15: Correlacions dels mètodes de la figura 5.29 detallades per clúster. En negreta es destaquen les correlacions superiors a 0.5.

A nivell de clúster s'observa que el clúster 2 és el que presenta més diferència entre els pesos obtinguts mitjançant l'ajust automàtic i els pesos obtinguts perceptivament obtenint una correlació entre ells aproximadament del 0.21. En canvi, el clúster 3 presenta una major similitud entre els pesos obtinguts automàticament i els pesos evolucionats per l'aiGA obtenint una correlació entre ells de 0.51. En aquest sentit es pot deduir que l'ajust mitjançant MLR/NNLS resulta perceptivament més adequat en uns contextos específics (clúster 3) respecte d'altres (clúster 2).

Addicionalment, a la figura 5.30 es pot observar la comparativa dels pesos extrets mitjançant mediana respecte als valors obtinguts segons els models latents. En aquest cas, és important observar que els models latents poden alterar l'ordre d'importància dels pesos dins del patró. Com es pot observar, el clúster que presenta més discrepància segons les diferents metodologies perceptives és el clúster 2 ($corr \approx 0.92$) on es pot veure que els mètodes latents provoquen una forta correcció a w_5 =PIT.C. Convé recordar que el clúster 2 és un dels dos clústers (juntament amb el clúster 3) que presentava un índex de convergència intrausuari (ρ) més baix, permetent intuir que existeix més d'una solució bona en el problema, fet que queda confirmat amb la l'alta discrepància entre Median, SOM i GTM. Si es continua analitzant amb detall el clúster 2, es pot veure que Median considera el pes més important w_9 =PosInSyl.L, mentre que SOM considera que és w_5 =PIT.C i



(a) Clúster 1.

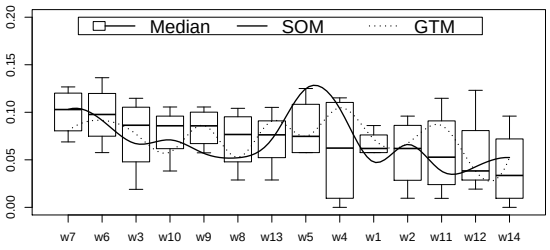
(b) Clúster 2.

(c) Clúster 3.

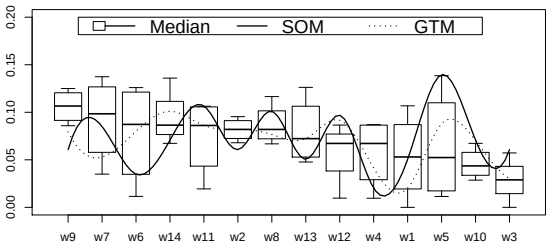
(d) Clúster 4.

(e) Clúster 5.

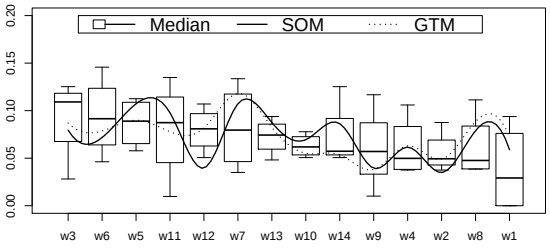
Figura 5.29: Patrons de pesos consensuats segons les diferents metodologies considerades (MLR/NNLS, Median, SOM, GTM) on w_1 =PIT.T, w_2 =ENE.T, w_3 =DURL.T, w_4 =DURR.T, w_5 =PIT.C, w_6 =ENE.C, w_7 =MFC.C, w_8 =PosInEG.L, w_9 =PosInSyl.L, w_{10} =PosInWord.L, w_{11} =Prev.L, w_{12} =Next.L, w_{13} =POS.L, w_{14} =Stress.L.



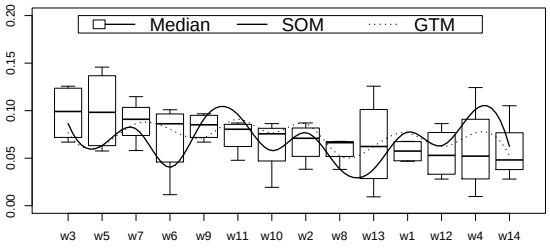
(a) Clúster 1.



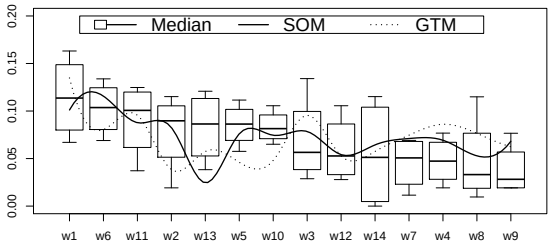
(b) Clúster 2.



(c) Clúster 3.



(d) Clúster 4.



(e) Clúster 5.

Figura 5.30: Clústers de pesos consensuats mitjançant models latents comparats amb els patrons obtinguts pel consens de mediana. Les línies ajunten els valors centrals de cada distribució. Els pesos s'ordenen de major a menor segons mediana on w_1 =PIT.T, w_2 =ENE.T, w_3 =DURL.T, w_4 =DURR.T, w_5 =PIT.C, w_6 =ENE.C, w_7 =MFC.C, w_8 =PosInEG.L, w_9 =PosInSyl.L, w_{10} =PosInWord.L, w_{11} =Prev.L, w_{12} =Next.L, w_{13} =POS.L, w_{14} =Stress.L.

GTM considera w_{14} =Stress.L. La discrepància del clúster 5 ($corr \approx 0.93$) confirma també els baixos indicadors de l'aiGA en aquest clúster.

En termes generals, les correlacions entre els diferents mètodes perceptius són: $corr$(Median,SOM) ≈ 0.93, $corr$(Median,GTM) ≈ 0.95 i $corr$(SOM,GTM) ≈ 0.95, essent els mètodes Median i SOM els que presenten més discrepància entre sí.

L'estudi presentat permet demostrar que els models latents confirmen la rellevància dels indicadors per determinar la bondat de les solucions obtingudes, com s'ha vist per exemple en el cas del clúster 2, on l'índex de convergència (ρ) intra-usuari baix s'ha convertit en correccions fortes per part dels models latents. A més, respecte la discussió de l'estat de l'art sobre l'idoneïtat d'emprar subcostos lingüístics o acústics a la funció de cost, els resultats demostren que no hi ha una competència clara entre ells permetent cooperar de manera conjunta per obtenir bones solucions finals.

5.8 Validació de l'ajust perceptiu dels pesos usant aiGA

En aquest apartat es detallen les proves perceptives per a validar les diferents propostes que s'han presentat en aquest capítol. Alhora, es realitza una nova validació que permet consolidar la metodologia d'ajust mitjançant aiGA respecte altres tècniques d'ajust de pesos de l'estat de l'art (MLR/NNLS i MOS-*Postmapping*) en un escenari real d'ajust de pesos.

Primerament, a l'apartat 5.8.1 s'expliquen els ajustos preliminars que estudien diferents aspectes: per una banda, els models latents de consens dels pesos i, per altra, s'estudia la vigència de les divisions contextuals (lingüístiques i fonètiques) en els pesos obtinguts de manera perceptiva . Posteriorment, l'apartat 5.8.2 explica amb detall l'obtenció de pesos perceptius mitjançant MOS-*Postmapping*. Finalment, a l'apartat 5.8.3 es compara perceptualment l'aiGA respecte les altres tècniques representatives de l'estat de l'art: MLR/NNLS i MOS-*Postmapping*. En aquesta última comparació s'estudien dos aspectes addicionals: el primer aspecte que s'estudia és la idoneïtat de la funció de cost clàssica de (Hunt i Black, 1996) basada en una distància de Manhattan ponderada respecte la funció de cost RMS (Toda *et al.*, 2006) basada en una distància euclídea. El segon aspecte que s'estudia és la diferència perceptiva que hi ha entre treballar amb els pesos a nivell global (una combinació per tot el corpus) i treballar amb pesos específics en funció dels diferents contextos (lingüístics i fonètics) que adopten les unitats.

En tot el capítol, les proves perceptives es realitzen emprant la metodologia *Mean Opinion Score* – MOS (ITU-T, 1996). El motiu per escollir MOS per avaluar la qualitat sintètica

és que permet disposar d'índexs de qualitat absoluts (i no comparatius) dels diferents sistemes. Tanmateix, cal recordar que alguns investigadors (Alvarez i Huckvale, 2002; Sityaev *et al.*, 2006) sostenen que les comparacions directes per parelles (proves de preferència duals) són millors que les valoracions MOS globals, en termes de discriminació entre sistemes similars. En aquest sentit, s'ha adaptat la metodologia MOS clàssica a una presentació paral·lela de dos estímuls diferents a l'usuari per així poder obtenir els avantatges dels dos mètodes. En altres paraules, es presenten a l'usuari dos estímuls diferents comparables entre sí (és a dir la mateixa frase sintetitzada amb dos pesos diferents) i l'usuari avalua cadascuna d'elles independentment. Així, al avaluar-les en paral·lel, indirectament estableix una gradació comparativa entre elles (puntuació A - puntuació B).

Addicionalment, a la presentació paral·lela s'hi mostra la mateixa frase amb veu natural enregistrada pel mateix locutor, per proporcionar una referència perceptiva a l'usuari. Per realitzar aquesta comparativa s'ha emprat la plataforma TRUE (Planet *et al.*, 2008), capaç de realitzar proves MOS, proves CMOS, o les dues simultàniament, entre d'altres.

5.8.1 Ajustos preliminars

GTM vs. SOM vs. Median

Un cop s'han obtingut els models aiGA per cada usuari s'han de consensuar els pesos que millor representin les evolucions que han realitzat els diferents usuaris. En l'apartat 5.6.1 s'han descrit els models latents respecte la metodologia basada en mediana emprada fins al moment. Posteriorment, en l'apartat 5.7.2 s'ha explicat com variaven els pesos dins cada clúster en funció de la tècnica de consens emprada. Tanmateix queda pendent analitzar el comportament dels mètodes latents en termes de qualitat de les síntesis finals.

Per comparar la idoneïtat dels diferents mètodes latents per tal d'aconseguir la màxima qualitat sintètica, es realitza una prova peceptiva on es comparen els pesos obtinguts a nivell de clúster mitjançant les tres metodologies de consens de pesos esmentades (Mediana, SOM i GTM). La prova consisteix en un MOS paral·lel que compara per parelles les 3 configuracions a través de 10 frases escollides aleatòriament del corpus. La prova la realitzen un total de 33 usuaris, que inclouen els 8 usuaris (experts en tecnologies de la parla) que han realitzat l'ajust dels pesos mitjançant aiGA i 25 usuaris que no tenen expertesa en síntesi de veu.

A la figura 5.31 s'observen els resultats perceptius obtinguts. Es pot observar una lleugera superioritat del GTM respecte el consens per Mediana i el consens mitjançant SOM. Concretament, les puntuacions mitjanes que s'obtenen són $MOS_{GTM} = 3.50$, $MOS_{SOM} =$

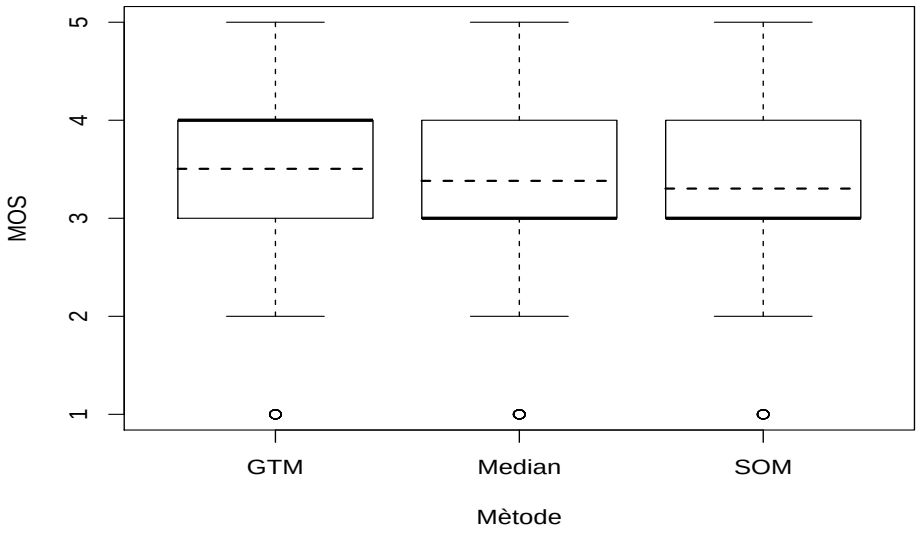


Figura 5.31: Comparativa MOS dels diferents models latents emprats (la línia de punts mostra el valor corresponent a la mitjana).

3.30 i $MOS_{Mediana} = 3.38$. Si s'estudia la significança estadística segons la prova t de Student s'obté que GTM és millor que Mediana ($p = 0.03$), que GTM és millor que SOM ($p = 7.1 \cdot 10^{-5}$) i que SOM no és significativament millor que Mediana ($p = 0.29$). Aquestes dades indiquen que la metodologia mitjançant GTM és el millor mètode per consensuar els pesos dels diferents usuaris respecte SOM i Mediana en termes de qualitat sintètica obtinguda.

Tanmateix, per ratificar aquests resultats s'analitzen les puntuacions obtingudes mitjançant la prova de comparativa directa (CMOS) juntament amb les significances obtingudes a través de la prova de signe de Wilcoxon (Hollander i Wolfe, 1973). En aquest cas s'obté que GTM $>$ Median ($p = 0.012$), GTM $>$ SOM ($p = 0.007$). A més, s'obté que Median $>$ SOM ($p = 0.009$), dada que no es podia obtenir mitjançant l'avaluació MOS normal.

CepstralTree vs. *aiGATree*

Un cop contrastada la validesa dels mètodes latents, queda per validar l'arbre de decisió a emprar en el sistema de síntesi final.

La diferència dels valors obtinguts entre els mètodes d'ajust automàtic i els mètodes d'ajust perceptius posa en dubte la vigència de les preguntes contextuals (fonètiques i lingüístiques) que identifiquen els diferents patrons. Cal recordar que el conjunt de preguntes, l'arbre pròpiament dit, s'ha obtingut amb els pesos entrenats automàticament.

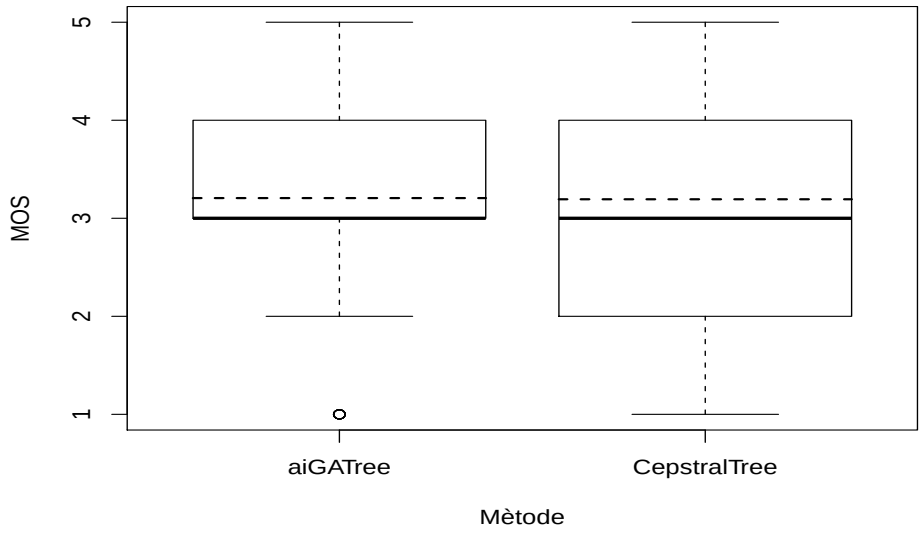


Figura 5.32: Comparativa MOS dels diferents arbres emprats.

Com a alternativa, es proposa reconstruir l'arbre de decisió (reformular les preguntes i per tant els grups) amb els pesos perceptius obtinguts. A més, en el nou arbre el nombre de patrons no resulta crític ja que es poden obtenir tants grups (nombre de pesos) com sigui necessari en termes d'augmentar la precisió. Així, els pesos obtinguts de manera perceptiva es poden associar amb els diferents contextos fonètics i lingüístics.

En aquest sentit, es comparen les síntesis generades a partir dels patrons de pesos obtinguts segons els dos arbres de decisió esmentats. En les proves, l'arbre de decisió original obtingut amb els pesos ajustats automàticament s'identifica com *cepstralTree* i l'arbre de decisió construït amb els pesos ajustats amb l'aiGA s'identifica com *aiGATree*. La prova torna a consistir en 10 frases escollides aleatòriament i els pesos perceptius obtinguts mitjançant aiGA. Els usuaris són els mateixos que en la prova anterior: 33 usuaris que inclouen els 8 experts en síntesi i 25 usuaris sense expertesa en síntesi de veu.

Els resultats obtinguts es mostren a la figura 5.32. En aquest cas s'accentua la similitud entre els dos mètodes. Concretament s'obté que $MOS_{cepstralTree} = 3.20$ i $MOS_{aiGATree} = 3.19$. La significància (t de Student) obtinguda entre ambdós mètodes és de $p = 0.87$ determinant així que els dos arbres són equivalents en termes de qualitat sintètica obtinguda.

A més, per poder constatar o desmentir aquesta equivalència s'analitzen la diferència de valors en comparació directa, anàlogament al mètode anterior. En aquest cas es confirma l'equivalència entre els dos mètodes ($p = 0.804$), fet que confirma que les preguntes contextuals (lingüístiques i fonètiques) emprades per identificar els diferents grups de pe-

sos segueixen sent vàlides després de l'ajust perceptiu amb aiGA.

5.8.2 MOS-*Postmapping*

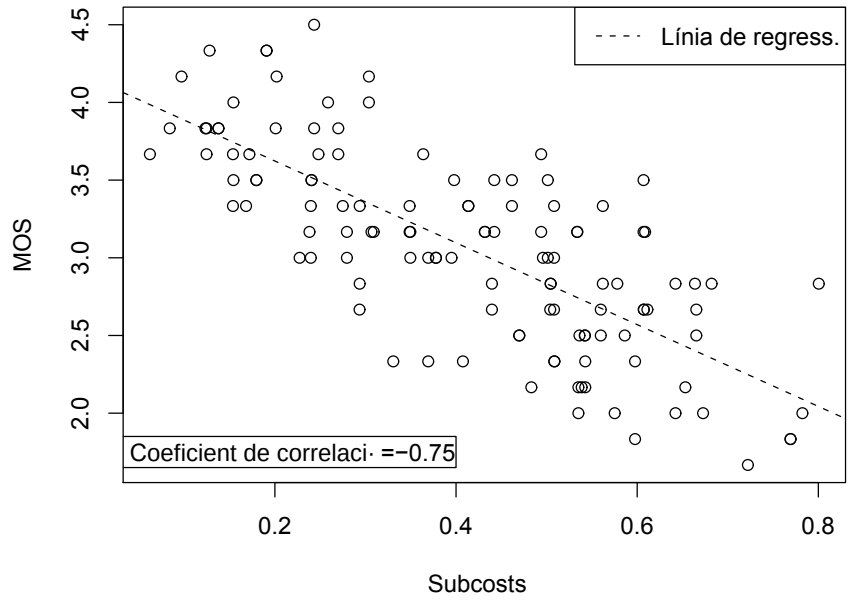


Figura 5.33: Regressió multilineal entre els subcostos amitjanats i les puntuacions MOS obtingudes a partir de la recopilació de preferències d'usuaris que avaluen diferents frases de test (veure apartat 4.7.2).

Donat que en aquest capítol es treballa amb un corpus més extens i amb més subcostos que en els experiments descrits en l'apartat 4.7, es vol tornar a comparar l'aiGA amb l'altra tècnica perceptiva de l'estat de l'art: MOS-*Postmapping*. D'aquesta manera, es pretén confirmar els resultats obtinguts comparant les diferents tècniques d'ajust perceptiu. En primer lloc, es realitzen unes proves perceptives per recollir les dades subjectives necessàries per a l'entrenament del MOS-*Postmapping*. En aquest sentit, es realitza una primera prova mitjançant MOS paral·lel que compara quatre configuracions diferents de la funció de cost. Aquestes configuracions combinen diferents pesos amb diferents variants de la mètrica d'integració dels subcostos dins de la funció de cost (veure apartat 2.3.5 i (Toda *et al.*, 2006)): *i*) aiGA + AVG, *ii*) aiGA + RMS, *iii*) MLR+AVG i *iv*) MLR+RMS. Els resultats d'aquestes proves es descriuen amb detall a l'apartat 5.8.3.

Les proves consten de 20 frases escollides aleatòriament obtenint així un valor MOS per cada síntesi realitzada en l'escala [1,5]. A diferència del capítol anterior (apartat 4.7.1)

no s'ha de realitzar cap transformació de les puntuacions (CMOS → MOS) donat que en la prova MOS paral·lel, les puntuacions de cada frase s'obtenen directament en l'escala [1,5].

Llavors, tal com s'ha explicat en l'apartat 4.7.1, s'obtenen els pesos globals que millor mapen els subcostos de cada frase segons les puntuacions MOS mitjançant una regressió MLR/NNLS. El resultat de l'esmentada regressió es pot observar a la figura 5.33, on s'obté una correlació del -0.75, valor proper a les correlacions obtingudes per (Toda *et al.*, 2006) quan treballa en la globalitat de tot el corpus.

A més, la correlació obtinguda és superior a la correlació obtinguda en els experiments anteriors que era de -0.49 (veure la figura 4.21). En l'apartat de discussió 5.8.4 s'analitzen els motius d'aquesta millora.

Els valors de pesos obtinguts en l'ajust per MOS-*Postmapping* són els següents: Stress.L= 0.25, ENE.C= 0.17, PosInEG.L= 0.13, Prev.L= 0.10, PosInSyl.L= 0.09, ENE.T= 0.08, PosInWord.L= 0.07, POS.L= 0.06, PIT.T= 0.04, MFC.C= 0.02, DURL.T= 0.00, DURR.T= 0.00, Next.L= 0.00 i PIT.C= 0.00.

5.8.3 Experiments i Resultats

Un cop s'han realitzat els ajustos preliminars per obtenir la configuració de l'aiGA (arbre original amb pesos consensuats segons GTM) i els pesos segons MOS-*Postmapping*, s'ha de confirmar la seva idoneïtat respecte els altres mètodes d'ajust siguin aquests perceptius (subjectius) o objectius.

Aquesta validació es divideixen en dues etapes: *i*) estudiar la idoneïtat de l'aiGA respecte la metodologia automàtica (MLR/NNLS), prova que s'aprofita per avaluar la distància de la funció de cost (Manhattan, euclídea...) i *ii*) confirmar la bondat de l'aiGA respecte el *MOS-Postmapping*, comprovant alhora, la idoneïtat de l'ajust de pesos a nivell de clúster respecte l'ajust de pesos a nivell global (veure apartat 5.4.1) dels pesos.

MLR vs. aiGA / RMS vs. AVG

La primera prova estudia la validesa de l'aiGA respecte la metodologia automàtica MLR/NNLS a nivell perceptiu, i combina les dues distàncies típicament emprades en la funció de cost (Manhattan vs. Euclídea) (Toda *et al.*, 2006). En aquesta prova es descarten els pesos ajustats mitjançant GA. El motiu ve donat pel fet que l'esmentada validació ja s'ha realitzat prèviament mitjançant un corpus semblant (veure figura 5.15).

La incorporació en l'anàlisi de l'estudi de la distància de la funció de cost obeeix al

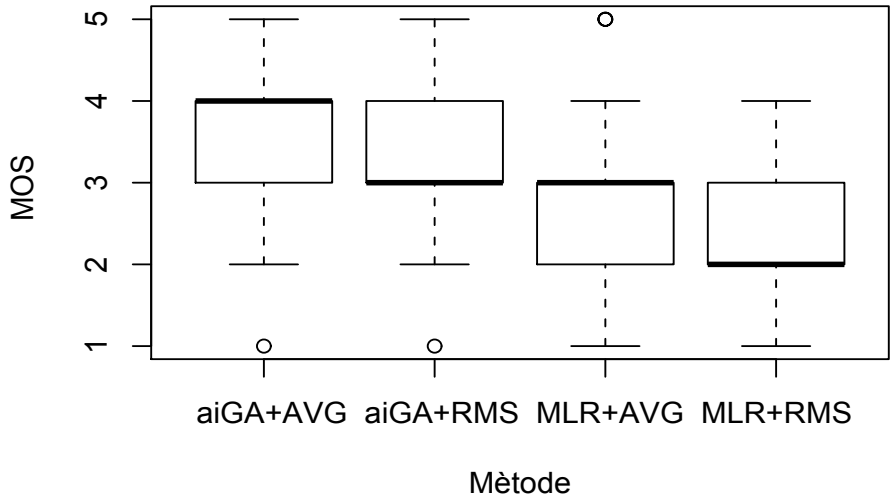


Figura 5.34: Comparativa MOS de l'aiGA (pesos a nivell clúster) respecte MLR/NNLS (pesos a nivell d'unitat) combinant diferents mètriques d'integració de subcostos a la funció de cost (avg=Manhattan / rms=euclídea) pel corpus *uvig_dav_es*.

fet de considerar la millor distància per al MOS-*Postmapping*. Concretament es compara la mètrica clàssica de (Hunt i Black, 1996) (anomenada *average* i referida com a AVG) i la mètrica RMS proposada per (Toda *et al.*, 2006). Per tant, a la primera prova es comparen quatre sistemes diferents (aiGA+AVG, aiGA+RMS, MLR+AVG i MLR+RMS) que combinen dues metodologies d'ajust automàtic de pesos (aiGA i MLR/NNLS) i dues distàncies de la funció de cost (AVG i RMS).

La informació obtinguda en aquesta validació és la que s'ha emprat per obtenir els pesos segons MOS-*Postmapping* (veure apartat 5.8.2). La prova la varen realitzar 7 dels 8 usuaris que prèviament havien ajustat els pesos mitjançant aiGA. La prova es tracta d'un MOS paral·lel que combina les 4 configuracions diferents a través de 20 frases escollides aleatòriament del corpus.

Els resultats obtinguts es mostren a la figura 5.34. Les dades evidencien una clara superioritat del mètode perceptiu (aiGA) respecte els mètodes de regressió automàtics (MLR/NNLS). Les puntuacions que s'obtenen són: $MOS_{aiGA\_avg} = 3.539$, $MOS_{aiGA\_rms} = 3.411$, $MOS_{MLR\_avg} = 2.944$ i $MOS_{MLR\_rms} = 2.361$. A l'analitzar les significança de les diferències mitjançant la t de Student amb correcció de Bonferroni s'obté que aiGA_avg és equivalent a aiGA_rms ($p = 0.743$) essent la resta de diferències significants ($p < 0.001$).

Tanmateix, malgrat que només hi ha igualtat entre aiGA_avg i aiGA_rms es procedeix a analitzar la comparació usant CMOS per obtenir les diferències directes dels mètodes. En

aquest cas, es confirma la igualtat entre aiGA_avg i aiGA_rms ($p = 0.105$) a la vegada que es confirmen la resta de diferències ($p < 0.001$). Les significàncies del CMOS es calculen mitjançant la prova de signe de Wilcoxon (Hollander i Wolfe, 1973).

Aquest resultats manifesten, per una banda, la clara superioritat de l'ajust de pesos perceptiu respecte l'ajust mitjançant MLR/NNLS i distàncies cepstrals. Addicionalment, els resultats indiquen que la mètrica d'integració de la funció de cost (RMS / AVG) no resulta tant determinant com la ponderació dels subcostos a l'hora d'obtenir una síntesi d'alta qualitat. No obstant això, quan la ponderació és dolenta (MLR/NNLS) si que és capaç de determinar diferències prioritzant, en els 14 subcostos analitzats i pel corpus *uvig_dav_es*, la distància clàssica de Manhattan (Hunt i Black, 1996).

MOS vs. *aiGA Global* vs. *aiGA Clustered*

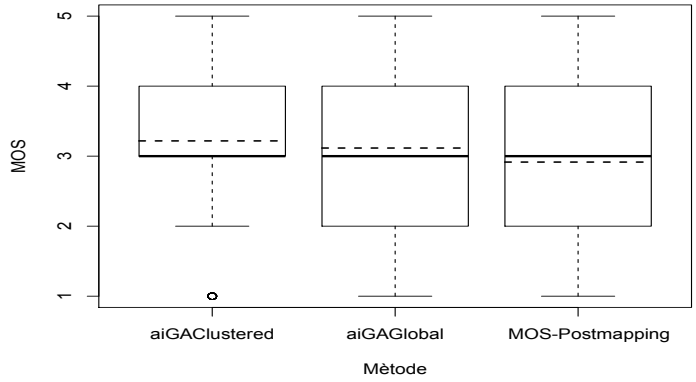


Figura 5.35: Comparativa MOS dels diferents mètodes perceptius d'ajust de pesos estudiats.

Un cop s'ha validat la millora que presenta la metodologia d'ajust de pesos basada en l'aiGA respecte l'ajust automàtic (MLR/NNLS) per les dues distàncies d'integració avaluades (AVG i RMS), queda pendent validar l'aiGA respecte MOS-*Postmapping*. Una de les principals diferències del MOS-*Postmapping* respecte la metodologia proposada (aiGA) es troba en el nivell d'ajust de pesos. Mentre que el MOS-*Postmapping* ajusta una sola combinació de pesos per a tot el corpus, s'usa l'aiGA per a trobar diferents patrons de pesos en funció dels diferents contexts fonètics i lingüístics en els que es troba la unitat a sintetitzar (veure apartat 5.4.1). En altres paraules, es treballa amb l'aiGA amb més nivell de precisió que el MOS-*Postmapping*, cosa que comporta que la comparació entre els dos mètodes no es realitzi en igualtat de condicions pel que fa a el nivell de precisió de l'ajust. No obstant això, es en futurs estudis es podria analitzar el MOS-*Postmapping* a nivell de clúster.

Per tal de realitzar una comparativa justa, a partir dels diferents pesos obtinguts mitjançant aiGA s'extreu un patró de pesos global únic que guiarà la síntesi i permetrà estudiar dos aspectes: *i*) la millora de l'ajust a nivell de clúster respecte l'ajust a nivell global i *ii*) obtenir una comparativa aiGA i MOS-*Postmapping* considerant el mateix nivell d'ajust. L'aiGA a nivell de clúster s'identifica en les proves com a *aiGAClustered*, l'aiGA amb un patró de pesos global per totes les unitats s'identifica com *aiGAGlobal* i el MOS-*Postmapping* resta de la mateixa forma.

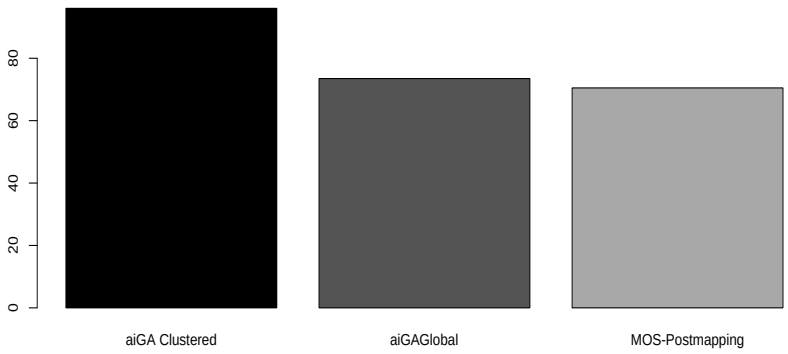
Les proves de validació es realitzen simultàniament a les proves preliminars d'ajust explicades a l'apartat 5.8.1. El motiu és poder garantir el mateix nombre d'usuaris a totes les proves. Per tant, la validació entre les tres configuracions compta amb la participació de 33 usuaris (inclosos els 8 usuaris que havien realitzat l'ajust de pesos mitjançant aiGA).

A la figura 5.35 es mostren els resultats obtinguts considerant les preferències dels usuaris segons l'escala MOS. En aquest cas, i a diferència de la comparació amb MLR/NNLS, les diferències entre les metodologies perceptives resulten més discretes. Les puntuacions que s'obtenen són $MOS_{aiGAClustered} = 3.21$, $MOS_{aiGAGlobal} = 3.11$ i $MOS_{MOS-Postmapping} = 2.91$. Si es comprova la significància d'aquestes diferències mitjançant la prova t de Student amb correcció de Bonferroni s'obté que *aGAClustered* i *aiGAGlobal* són equivalents ($p = 0.18$), *aiGAClustered* és millor que MOS-*Postmapping* ($p = 7.3 \cdot 10^{-8}$) i *aiGAGlobal* resulta millor que MOS-*Postmapping* ($p = 6.3 \cdot 10^{-4}$).

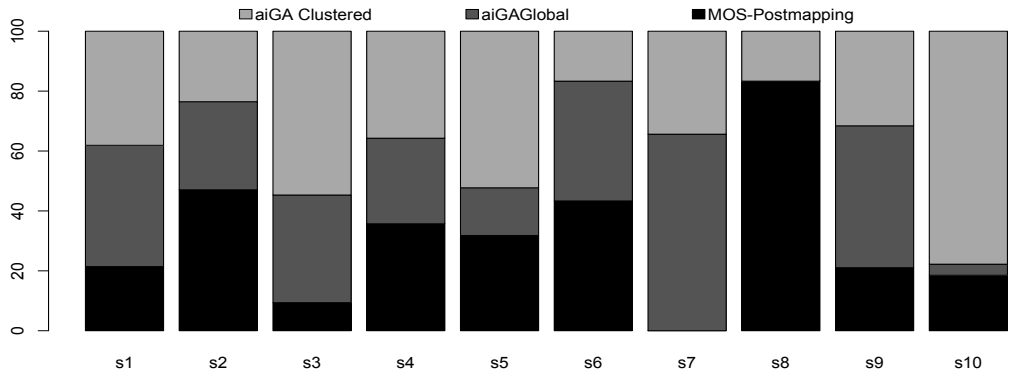
En termes de comparació directa de les diferents configuracions (via CMOS) s'aprecien diferències que no s'han pogut distingir mitjançant les puntuacions absolutes. Concretament s'obté que $aiGAClustered$ és millor que $aiGAGlobal$ ($p = 5.02 \cdot 10^{-3}$), $aiGAClustered$ és millor que MOS-*Postmapping* ($p = 2.172 \cdot 10^{-5}$) i $aiGAGlobal$ és millor que MOS-*Postmapping* ($p = 6.133 \cdot 10^{-5}$).

A part d'analitzar els resultats de preferència absoluta i relativa, els resultats també s'estudien considerant la millor configuració per cada dupla <usuari,frase>. Aquesta anàlisi permet comparar els resultats amb les anàlisis realitzades en les proves de viabilitat i exposades en els apartats 3.4.8 i 4.7.2. En aquest sentit, s'assigna un vot a la configuració guanyadora de cada dupla <usuari,frase>. Els resultats obtinguts es mostren a la figura 5.36. En termes absoluts (figura 5.36(a)) es pot observar una clara preferència per la metodologia *aiGAClustered* a l'obtenir la preferència de 96 duples <usuari,frase>. Es tracta d'un nombre de vots clarament superior a *aiGAGlobal* que es prefereixen de 74 duples i MOS-*Postmapping* que és el preferit de 71 duples.

Tanmateix, a la figura 5.36(b) es pot observar que la configuració guanyadora depèn de la frase que s'empra per realitzar la síntesi. Per exemple, la frase s8 denota una clara



(a) Distribució dels mètodes d'ajust guanyadors per usuari de la validació perceptiva. L'eix vertical denota el nombre de vots total que adopta cada configuració



(b) Distribució dels mètodes d'ajust guanyadors per usuari de la validació perceptiva detallat per a totes les frases.

Figura 5.36: Comparativa perceptiva entre els diferents pesos segons les contribucions d'aquest capítol pel corpus *uvig_dav_es*.

superioritat del MOS-*Postmapping* respecte l'*aiGAClustered* mentre que la frase s10 denota una clara superioritat de l'*aiGAClustered* respecte les altres configuracions.

5.8.4 Discussió

En aquest apartat s'han validat de manera perceptiva diferents aspectes de l'ajust de pesos mitjançant aiGA. Aquestes validacions s'han dividit en dues etapes: una d'ajustos preliminars que valorava les tècniques emprades per obtenir els pesos finals del sistema a partir dels models aiGA i una altra de validacions externes que valorava l'aiGA respecte els altres mètodes d'ajust de pesos considerats: MLR/NNLS (objectiu) i MOS-*Postmapping* (subjectiu). En aquesta etapa de validacions externes s'ha analizat alhora l'impacte perceptiu de les distàncies d'integració (AVG, RMS) dels diferents subcostos en la funció de cost, alhora que es generalitzaven els pesos de l'aiGA a nivell global (un vector per tot el corpus) per fer-lo competir amb MOS-*Postmapping* amb les mateixes condicions en termes de precisió de pesos (sense que aquests depenguéssin del context lingüístic i fonètic de la unitat a sintetitzar). A continuació es discuteixen els aspectes més rellevants de les proves realitzades.

Ajustos preliminars

Respecte l'anàlisi de la qualitat sintètica dels pesos consensuats mitjançant models latents, el GTM adopta un millor comportament que SOM i que Median. Per explicar aquest fet, a l'apartat 5.6.3 s'ha descrit com el GTM és una tècnica dissenyada específicament per obtenir un model robust i superar les mancances del SOM, que no deixa de ser un mètode no determinista (heurístic). De fet, el tret diferencial del GTM és deguda a l'adaptació a l'espai de dades mitjançant la fase d'esperança-maximització (EM), que el fa robust respecte el soroll que hi pugui haver-hi en les dades d'entrenament.

En canvi, es constata la gran diferència que hi ha entre els pesos ajustats automàticament (MLR/NNLS) i els pesos obtinguts de manera perceptiva (corr=0.38). Aquesta diferència qüestiona l'arbre de decisió que s'empra per obtenir els patrons de pesos. Com a alternativa, s'ha proposat construir un nou arbre de decisió a partir dels pesos obtinguts amb l'aiGA per veure si els patrons segueixen essent vàlids. Quan els dos arbres de decisió són comparats de manera perceptiva, es conclou que no hi ha diferències notables en termes de qualitat sintètica, permetent concloure que els patrons obtinguts amb els pesos automàtics segueixen essent vàlids un cop s'han obtingut els pesos de manera perceptual (aiGA). En altres paraules, l'especificitat marcada pels contextos lingüístics i fonètics, que

permet escollir quins pesos s'han d'emprar a la funció de cost, segueix essent vàlida per pesos obtinguts de manera perceptiva, encara que aquests adoptin uns valors diferents respecte els pesos que han permès definir els patrons.

MOS-*Postmapping*

Quan s'analitzaven els valors de pesos obtinguts mitjançant MOS-*Postmapping*, s'observa que el MOS-*Postmapping* prioritza els subcostos lingüístics amb excepció de les energies (ENE.C i ENE.T), fet que constata que els subcostos lingüístics poden cooperar (i no competir) amb els subcostos acústics. A més cal destacar que l'ordre obtingut guarda certa relació amb els pesos ajustats a mà per Clark *et al.* (2007) i presentats a la taula 5.11. En ambdúes ponderacions (MOS-*Postmapping* i manual) PosInEG.L i Stress.L adopten els valors alts de pesos deixant a un nivell intermedi PosInWord.L i POS.L i una ponderació baixa per Next.L. Les diferències es troben en Prev.L i PosInSyl.L que adopten un valor alt en MOS-*Postmapping* a diferència del seu ajust heurístic a mà. Per tant, hi ha coincidencia en 5 dels 7 subcostos lingüístics analitzats. Convé recordar que el propi treball de Clark *et al.* (2007) sosté que l'ajust manual per part d'un expert resulta suficient per obtenir una bona qualitat sintètica. La semblança d'ambdues ponderacions (manual i MOS-*Postmapping*) constata que no fa falta tenir coneixement dels subcostos que intervenen en la síntesi (coneixement expert) per realitzar la mateixa ponderació que realitzaria un expert a base de prova i error.

A més, cal destacar que pel fet de disposar d'un corpus més extens (8 min. *url_fer_ct*→ 1.9h *uvig_dav_es*) i amb més subcostos (7 → 14) s'ha conseguit millorar la correlació obtinguda mitjançant aquest mètode (MOS-*Postmapping*), passant d'una correlació de $r = -0.49$ (*url_fer_ct*) a una correlació de $r = -0.75$ (*uvig_dav_es*), valor proper a les correlacions obtingudes per (Toda *et al.*, 2006) quan treballa en la globalitat de tot el corpus.

Validacions externes

En aquest apartat, s'ha contrastat de nou la superioritat de la metodologia basada en l'aiGA respecte els mètodes d'ajust de pesos MLR/NNLS basat en distàncies cepstrals, dins l'escenari d'ajust per subunitat contextualitzada. A la vegada, s'han comparat les distàncies d'integració de subcostos de la funció de cost comparant la distància de Hamming (AVG) amb la distància euclídea (RMS). En aquest sentit s'ha observat que si la ponderació de pesos és bona, l'impacte de la distància d'integració resulta gairebé nul. En canvi, en ponderacions de pesos dolentes o deficients (MLR/NNLS), sí que s'hi poden

apreciar diferències, però a favor d'AVG en el cas dels 14 subcostos emprats (7 acústics i 7 lingüístics) per el corpus *uvig_dav_es*.

Aquests resultats són diferents dels obtinguts per (Toda *et al.*, 2006) on la mètrica RMS era la que correlava lleugerament millor amb les preferències dels usuaris. Tanmateix, existeixen diferents aspectes que s'han de tenir en compte en aquesta comparativa: el primer aspecte és que Toda *et al.* (2006) no realitzava comparació directa de les dues distàncies mitjançant MOS o CMOS. El seu estudi es basa en el càlcul del coeficient de correlació que obté el model de regressió entre els subcostos i les puntuacions MOS havent emprat AVG per a totes les síntesis. El segon aspecte és que la diferència d'aquests coeficients de correlació, segons Toda *et al.* (2006), és relativament petita ($corr(\text{AVG}) = -0.808$, $corr(\text{RMS}) = -0.840$), tot i que és significativa. El tercer aspecte a considerar és que Toda *et al.* (2006) no treballa segons l'esquema clàssic de (Hunt i Black, 1996) quan realitza aquesta comparativa sinó que realitza la selecció d'unitats amb subcostos basats en prosòdia i seqüències Mel-cepstrals de la veu original, sense tenir en compte subcostos lingüístics o fonètics (de naturalesa discreta). Tanmateix, a posteriori, sí que realitza un estudi no perceptiu de l'impacte de la mètrica RMS en els subcostos segons l'esquema clàssic de (Hunt i Black, 1996) combinant diferents subcostos (acústics, lingüístics i fonètics). No obstant això, no valida els efectes d'aquest impacte de manera perceptiva limitant el seu estudi a una aproximació analítica (i no subjectiva) al problema.

Respecte la comparativa amb els altres mètodes perceptius, s'aprecien dues evidències destacables. La primera evidència és que el sistema de síntesi que selecciona els pesos en funció del context fonètic i lingüístic obté una millor acceptació per part dels usuaris que el sistema que empra la mateixa combinació de pesos per a tot el corpus, encara que aquests s'obtinguin mitjançant aiGA. La segona evidència és la preferència cap als pesos obtinguts mitjançant aiGA respecte els pesos subjectius obtinguts mitjançant regressió lineal (MOS-*Postmapping*), fins i tot quan els dos mètodes perceptius treballen a nivell global (configuració única de pesos per a tot el corpus). Aquesta evidència demostra que el problema d'ajust de pesos, tal com ja s'intuïa en l'apartat 3.2.4, és un problema d'ajust no lineal, almenys quan s'intenta solucionar globalment.

5.9 Conclusions

En aquest capítol s'ha realitzat un pas més dins l'esquema d'ajust perceptiu evolutiu presentat en el capítol 3. Fins al moment, l'ajust perceptiu evolutiu només s'havia validat conceptualment usant un corpus petit (*url_fet_ct* de 8 min) etiquetat a mà i amb una funció

de cost que només considerava set subcostos de la tipologia acústica.

Per reproduir un escenari d'ajust més real i d'acord amb l'estat de l'art ($> 1h$) en sistemes de síntesi de veu basats en selecció d'unitats (Taylor, 2009), s'ha canviat el marc de treball. Primerament s'ha canviat el corpus de veu a un corpus de veu dissenyat expressament per sistemes de síntesi de veu basats en selecció d'unitats (*uvig_dav_es* de 1.9h).

Un dels objectius plantejats en aquest capítol era millorar la fiabilitat dels mètodes automàtics en l'ajust de pesos mitjançant distàncies cepstrals. Quan s'estableixen patrons d'ajust de pesos a partir d'aquests pesos automàtics, es necessita que aquests siguin el més fidels possible a la informació espectral. Per assolir aquest objectiu es millora la normalització dels subcostos amb una nova funció de transformació (SQRT) i es treballa a un nou nivell d'ajust que té en compte la especificitat de la pròpia unitat marcada pels contextos fonètics i lingüístics que pren la pròpia unitat dins la funció de cost. En realitzar els ajustos a nivell de subunitat contextualitzada, s'aconsegueix optimitzar el modelat mitjançant regressió lineal (MLR/NNLS), millorant tant el coeficient de determinació R^2 com l'error quadràtic del modelat RMSE. Concretament es passa del $R^2_{UNIT} = 33.9\%$ i $RMSE_{UNIT} = 23.98$ a $R^2_{SUBUNIT} = 43.4\%$ i $RMSE_{SUBUNIT} = 9.307$. No obstant això, el GA no millora la seva fiabilitat en treballar a nivell de subunitat contextualitzada. Convé recordar que el disseny del GA (Alías i Llorà, 2003) assumeix, de manera no contrastada, que en l'amitjanat del cost total obtingut amb la ponderació segons els pesos avaluats dels subcostos de les 5 versions més properes cepstralment, s'obté el *fitness* dels pesos avaluats. Però en canvi no es té en compte que l'amitjanat no considera l'ordre entre elles. A part, el disseny del GA obeeix a un entorn sorollós provocat per seleccionar aleatòriament unitats diferents a l'hora d'avaluar les combinacions de pesos, premissa que no es respecta al treballar a nivell de subunitat. Per aquest motiu es deixa la porta oberta a seguir investigant en aquest sentit. De tota manera, es pot dir que s'aconsegueix una millora significativa de la fiabilitat dels pesos obtinguts mitjançant regressió lineal gràcies a les dues contribucions (millor normalitat i canvi de nivell d'ajust de pesos per subunitat) aconseguint així superar els objectius de precisió i robustesa plantejats en el transcurs de la tesi.

Per superar l'objectiu d'obtenir una metodologia robusta de detecció de patrons es realitza un estudi complet del procés d'agrupament (apartat 5.5) exposant prèviament les limitacions de realitzar el *clustering* de pesos mitjançant arbres de decisió de manera directa. Aquestes limitacions s'identifiquen com a problemes de *i*) *clustering* predictiu i *ii*) excés d'equilibrat. La seva conseqüència més directa és la incapacitat de l'arbre de decisió de trobar grups naturals en les dades si aquests no es poden obtenir de manera directa mitjançant una característica que defineix el context de la subunitat. Per superar aquestes

limitacions es proposa una tercera contribució que consta en separar la detecció de patrons en dues etapes: una etapa de *clustering* pròpiament dita i una etapa de classificació que associa cadascun dels patrons obtinguts segons els diferents contextos fonètics i lingüístics presents en el corpus. L'estudi (apartat 5.5.3), que considera varies distàncies de bondat de l'estat de l'art, determina que el *clustering* mitjançant esperança-maximització (EM) és la millor tècnica per trobar patrons en els pesos en la fase de *clustering* pròpiament dita. També determina, segons les mètriques de bondat aplicades, que el nombre ideal de patrons pel corpus estudiat és de 5. Llavors, es configura l'arbre de decisió per a que realitzi un bon mapatge dels diferents patrons a cada context sense veure's limitat pels problemes d'equilibrat.

En quart lloc en aquest capítol (apartat 5.6.1) es proposa assolir l'objectiu d'obtenir un model robust que mapi i consensuï les preferències dels diferents usuaris respecte els pesos escollits. En aquest sentit es proposa emprar models latents, amb gran acceptació dins les ciències socials per trobar patrons. Aquest fet permet obtenir un sol vector de pesos de manera que aquest representi el consens entre els diferents usuaris i per tant entre les diferents metodologies de consens aiGA plantejades. En concret, s'estudien dos models latents de l'estat de l'art: SOM i GTM. Els resultats demostren la superioritat de consensuar els pesos mitjançant GTM respecte Mediana o SOM: GTM és capaç d'obtenir una millor acceptació entre els usuaris. Validant així la hipòtesi que el model latent mitjançant GTM és apropiat per trobar un criteri consensuat entre els diferents usuaris.

El cinquè punt d'aquest capítol és l'estudi de la vigència de l'arbre de decisió cepstral un cop s'han ajustat els pesos de manera perceptiva. En aquest cas, s'estudia la viabilitat de reconstruir un l'arbre de decisió amb els pesos obtinguts perceptivament i veure si adopta diferències respecte l'original. Els resultats obtinguts demostren que els patrons originals (obtinguts amb pesos automàtics) segueixen essent vigents un cop s'ha realitzat l'ajust perceptiu, malgrat que els valors dels pesos que el conforma sigui diferent. En altres paraules, els diferents contextos lingüístics i fonètics que permeten escollir quins pesos cal emprar a l'hora de sintetitzar una unitat són vàlids abans i després de realitzar l'ajust de manera perceptiva.

El diagrama final de la metodologia d'ajust proposada en aquesta tesi es mostra a la figura 5.37 on es concreten les tècniques que presenten un millor comportament a cada etapa del procés.

Un cop s'han validat les diferents contribucions, resta pendent validar l'eficiència de l'aiGA per obtenir bons pesos de manera perceptiva en un entorn real de selecció d'unitats que consideri subcostos de diferent naturalesa i un corpus extens. Paral·lelament, en

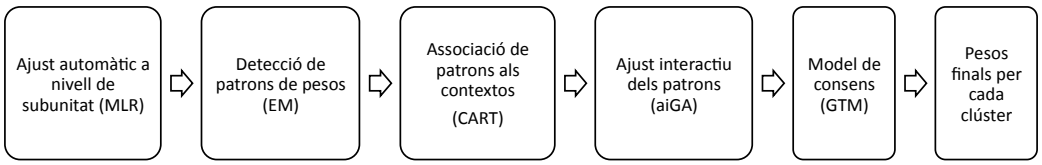


Figura 5.37: Diagrama de les etapes de la metodologia proposada per ajustar pesos a nivell de clúster. Entre parèntesi es destaca la tècnica proposada en cada etapa.

aquest estudi s'hi incorpora l'estudi dos aspectes tangencials a la funció de cost que no havien estat avaluats. Primer s'estudia l'impacte de les diferents mètriques d'integració dels subcostos a la funció de cost, degut a que és un tema que afecta directament la funció de cost. A continuació, s'estudia l'impacte dels diferents nivells d'ajust de pesos en la qualitat final de la síntesi per determinar si una sola combinació de pesos per tot el corpus és suficient a nivell perceptiu o altrament és millor obtenir patrons de pesos en funció dels diferents contextos lingüístics i fonètics en els que es troben les subunitats del corpus. Els resultats obtinguts demostren l'eficiència de l'aiGA en un entorn real de selecció d'unitats respecte els altres mètodes d'ajust de pesos considerats, alhora que també demostren la idoneïtat de treballar amb vectors de pesos en funció dels diferents contextos lingüístics i fonètics que adopta la unitat. No obstant això, els resultats denoten, pels subcostos i corpus estudiat, que la distància aplicada (Hamming, euclídea, etc.) a la funció de cost per integrar dels diferents subcostos no afecta la qualitat de la síntesi si la ponderació dels subcostos és bona en termes d'acceptació perceptiva.

L'última conclusió que es pot manifestar a partir dels resultats obtinguts és que la millora en la fiabilitat dels pesos obtinguts mitjançant distàncies cepstrals (objectius) no aconsegueix reduir la gran diferència respecte els mètodes d'ajust perceptius en la qualitat final de la síntesi. La distància existent entre els pesos obtinguts mitjançant distàncies cepstrals i els pesos obtinguts mitjançant avaluació perceptiva deixa el camp obert per seguir investigant una mètrica espectral que s'adeqüi millor a les diferents percepcions de l'usuari, tal i com es detallarà en les conclusions d'aquesta tesi (veure capítol 6).