Treball Final de Màster

Reconeixement de veu mitjançant xarxes neuronals

Xavier Clos Calm

Màster en Tecnologies Aplicades de la Informació

Director/a: Jordi, Solé i Casals

Vic, Setembre de 2012

Universitat de Vic
Escola Politècnica Superior

REGISTRE D'ENTRADA

Núm.:……
DATA:……

Resum del Treball Final de Màster
Màster en Tecnologies Aplicades de la Informació

Títol: Reconeixement de veu mitjançant xarxes neuronals
Paraules clau: Xarxes Neuronals, Reconeixement de veu, Paraules aïllades, GMM, DTW, coeficients MFCC, Alfa6Uvic
Autor: Xavier Clos Calm
Direcció: Jordi Solé i Casals
Data: Setembre 2012

Resum

La interacció home-màquina per mitjà de la veu cobreix moltes àrees d'investigació. Es destaquen entre altres, el reconeixement de la parla, la síntesis i identificació de discurs, la verificació i identificació de locutor i l'activació per veu (ordres) de sistemes robòtics. Reconèixer la parla és natural i simple per a les persones, però és un treball complex per a les màquines, pel qual existeixen diverses metodologies i tècniques, entre elles les Xarxes Neuronals.

L'objectiu d'aquest treball és desenvolupar una eina en Matlab per al reconeixement i identificació de paraules pronunciades per un locutor, entre un conjunt de paraules possibles, i amb una bona fiabilitat dins d'uns marges preestablerts. El sistema és independent del locutor que pronuncia la paraula, és a dir, aquest locutor no haurà intervingut en el procés d'entrenament del sistema.

S'ha dissenyat una interfície que permet l'adquisició del senyal de veu i el seu processament mitjançant xarxes neuronals i altres tècniques. Adaptant una part de control al sistema, es podria utilitzar per donar ordres a un robot com l'Alfa6Uvic o qualsevol altre dispositiu.

Universitat de Vic
Escola Politècnica Superior

REGISTRE D'ENTRADA

Núm.:……
DATA:……

Final Project Summary
Master in Applied Information Technologies

Title: Speech recognition by artificial neural networks

Keywords: Neural Networks, Speech Recognition, Isolated Words, GMM, DTW, MFCC coefficients, Alfa6Uvic.

Author: Xavier Clos Calm

Director: Jordi Solé i Casals

Date: September 2012

Summary

The man-machine interaction by voice covers many areas of research. Among others, we have systems for speech recognition, speech synthesis and identification, verification and speaker identification, and voice activation (orders) for robotic systems. Recognizing speech is natural and simple for people, but it is a complex task for machines for which several methodologies and techniques exists, including artificial neural networks.

The aim of this work is to develop a Matlab based tool for recognition and identification of words spoken by a speaker, from a set of possible words and with good reliability within predetermined margins. The system is speaker independent, meaning that the user of the system has not been involved into the training process of the system.

During the project, an I/O interface has been designed that allows performing signal acquisition and processing of voice through neural networks and other techniques. Adapting a control part, the system could be used to give orders to a robot, for example Alfa6Uvic robot or other devices.

ÍNDEX

Índex de Figures

Índex de Taules

1. INTRODUCCIÓ

1. INTRODUCCIÓ

Bàsicament, el reconeixement de veu és un procés de classificació de patrons, que el seu objectiu és classificar el senyal d'entrada (ona acústica) en una seqüència de patrons prèviament apresos i emmagatzemats en uns diccionaris de models acústics i de llenguatge. Aquest procés de classificació suposa, que el senyal de veu pot ser analitzada en segments de curta duració i representar cada un dels segments pel seu contingut freqüencial, de forma semblant al funcionament de la oïda humana.

Al parlar de reconeixement de veu, ens podem imaginar innumerables camps d'aplicació; des de la domòtica, la intel·ligència artificial, ajuda a discapacitats físics, l'aviació tant civil com militar, les telecomunicacions i serveis afegits com la automatització dels serveis d'operadora, la validació de compres amb targeta de crèdit, etc.

Aplicacions com les mencionades han sigut per molt temps una fita pels investigadors, però malgrat els grans avanços tecnològics d'avui en dia, no ha sigut possible arribar a un resultat similar al de l'ésser humà. Lo que semblava ser un problema senzill, amb el temps s'ha convertit en una tasca cada vegada més complicada. La comprensió del que sentim no està delimitada solament a l'oïda, sinó que és una funció molt més complexa d'elaboració neurològica. L'oïda compleix la funció de captar el so, però el reconeixement del mateix és una funció purament cerebral.

1.1 Objectius

General

Aconseguir tenir una eina en Matlab que sigui capaç de reconèixer una paraula, pronunciada per un locutor, entre un conjunt de paraules possibles, amb una alta fiabilitat dins d'uns marges preestablerts. El sistema ha d'identificar la paraula d'un conjunt de paraules predeterminades o bé indicar que no és cap de les paraules del conjunt. A més, el funcionament ha de ser independent del locutor que digui la paraula. Es dissenyarà una interfície que permeti l'adquisició del senyal de veu i el seu processament mitjançant xarxes neuronals per tal de reconèixer la paraula pronunciada.

Específics

- Reconeixement de paraules aïllades
- De vocabulari restringit i limitat a unes poques paraules
- De locutor independent
- Elaborat amb Matlab
- Classificar mitjançant xarxes neuronals artificials
- Bon índex d'efectivitat (mínim un 80%)

1.2 Aplicació del reconeixedor de veu

Adaptant una part de control al sistema reconeixedor de veu d'aquest treball, és podria utilitzar per controlar el robot Alfa6UVic, donant determinades ordres verbals per qualsevol locutor ja que és un sistema independent del locutor.

En la *figura 1.1* podem veure un esquema complet del sistema en general

Figura 1.1 Esquema del sistema complet

La part de control queda fora de l'àmbit d'aquest treball, però es tractaria d'activar una sortida determinada en funció de la paraula reconeguda i mitjançant algun dispositiu de control com un PLC activar els motors adients dels braç robòtic

L'alfa6UVic *(fig.1.2)* és un braç robòtic de 6 Graus de llibertat dissenyat i desenvolupat a la Universitat de Vic, la seva funció principal és la utilització en docència, sobretot en les assignatures del grau Enginyeria Mecatrònica i el grau en Enginyeria Electrònica i Automàtica.

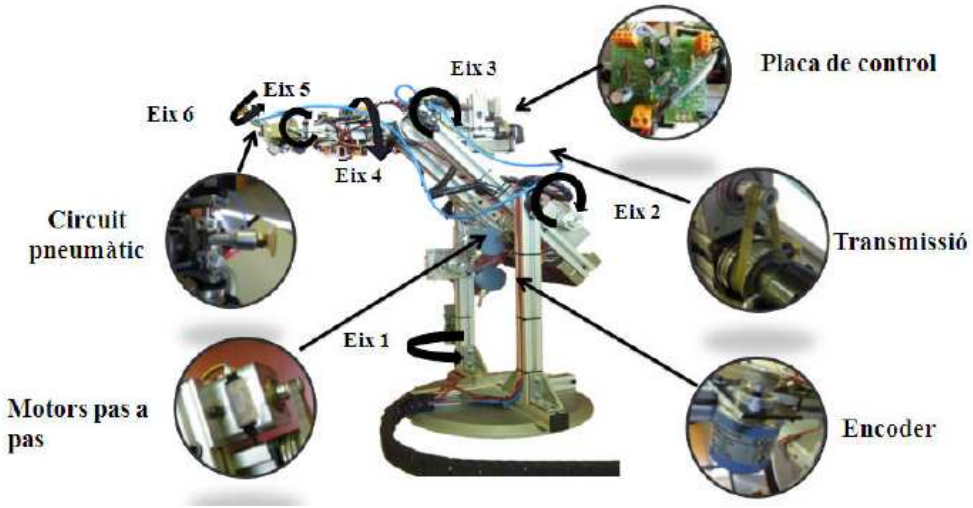


Figura 1.2 Robot Alfa6Uvic

El reconeixedor de veu s'ha entrenat per reconèixer el següent vocabulari:

"Dreta"	*"Esquerra"*
"Amunt"	*"Avall"*
"Gira Dreta"	*"Gira Esquerra"*
"Engega"	*"Para"*

Taula 1.1 Robot Alfa6Uvic

Amb aquestes paraules es podrien controlar tres eixos, més les funcions de parada i engegada del robot. Per controlar la resta d'eixos caldria afegir sis paraules o ordres més al sistema reconeixedor. Sense massa dificultat, això implicaria recollir noves mostres de veu i entrenar el sistema, que només caldria executar una vegada.

1.3 Metodologia

1 Recull d'informació bibliogràfica: Es va realitzar una recerca de diferents articles relacionats amb el reconeixement de veu, especialment els destinats a reconeixedors de paraules aïllades i independents de locutor.

2 Cerca d'un "toolbox" per Matlab de processament de veu

3 Creació del corpus de veus: Crear una base de dades enregistrant les paraules predeterminades de diversos locutors amb característiques diferents com l'edat i el gènere.

4 Pre-processament i Parametrització:

- Crear un "*script"* per aïllar el senyal de veu del silenci o soroll.
- Determinar quins són els paràmetres òptims en l'etapa de pre-processament i parametrització com per exemple el nombre de coeficients MFCC, el nombre de filtres òptim per calcular els coeficients, el millor tipus d'enfinestrament, etc.

5 MFCC-ANN:

- Proves d'efectivitat en la classificació, mitjançant Xarxes Neuronals amb els coeficients MFCC com a dades d'entrada.
- Obtenció dels millors paràmetres a aplicar a la Xarxa Neuronal fent proves empíriques de rendiment.

6 GMM:

- Obtenció dels histogrames de cada paraula a partir dels coeficients MFCC
- Proves d'efectivitat en la classificació mitjançant la generació de Models de mescles de Gaussianes.
- Es van provar diverses formes de generar els models, com un model per cada paraula o bé un model per cada coeficient.

- Proves de funcions que calculen les distàncies amb diferents mètriques entre els models i les paraula a reconèixer.

7 DTW : Proves d'efectivitat utilitzant l'algoritme DTW

8 DTW-ANN:

- Recerca de l'algoritme DTW més ràpid a Internet
- Generació d'una matriu d'entrenament composta per les distancies calculades amb el algoritme DTW entre els coeficients MFCC de senyals de referència i els MFCC de senyals d'entrenament. El mateix per
- Proves d'efectivitat utilitzant les distàncies calculades per l'algoritme DTW, com a dades d'entrada de diverses Xarxes Neuronals.

9 Desenvolupament de les interfícies gràfiques.

- Interfície de paràmetres
- Interfície de captura de mostres i referències
- Interfície d'entrenament i test
- Interfície del reconeixedor de veu real

1.4 Organització del treball

Capítol 2: Marc Teòric:

En aquest capítol s'intenta resumir els fonaments teòrics en que es basa el reconeixement de veu en general, començant per veure l'aparell fonedor i auditiu humà i com es pot modelar en un sistema automàtic.

Capítol 3: Marc Experimental:

En aquest capítol es descriuen els experiments realitzats a la pràctica mitjançant el Matlab, on es pot veure l'aplicació dels conceptes teòrics posats a la pràctica. Es descriuen les funcions de Matlab més importants.

També s'explica com es va portar a terme l'elaboració de la base de dades de veus o corpus de veus. Finalment es presenta la interfície gràfica dissenyada i els resultats obtinguts amb els diferents reconeixedors implementats.

Capítol 4: Conclusions:

S'exposen les conclusions del treball i quines son les possibles millores que es podrien aplicar al sistema desenvolupat.

5. Bibliografia:

Bibliografia utilitzada durant el procés d'elaboració del treball, sobretot abans de començar la part experimental.

6. Annex:

Tots els programes o funcions elaborats en Matlab i els obtinguts d'Internet.

2. MARC TEÒRIC

2. MARC TEÒRIC

2.1 Fisiologia de la veu

Per determinar les operacions d'un sistema automàtic de reconeixement de veu i locutor, és fonamental conèixer i determinar els mecanismes que han produït un missatge parlat, per després poder reproduir-los automàticament, de la mateixa manera que els mecanismes físics per descodificar el missatge. Per això es fa una petita introducció als conceptes fonamentals de la producció de parla, tant els òrgans físics *(figura 2.1-2.2)* com la producció del missatge.

2.1.1 Producció de la veu

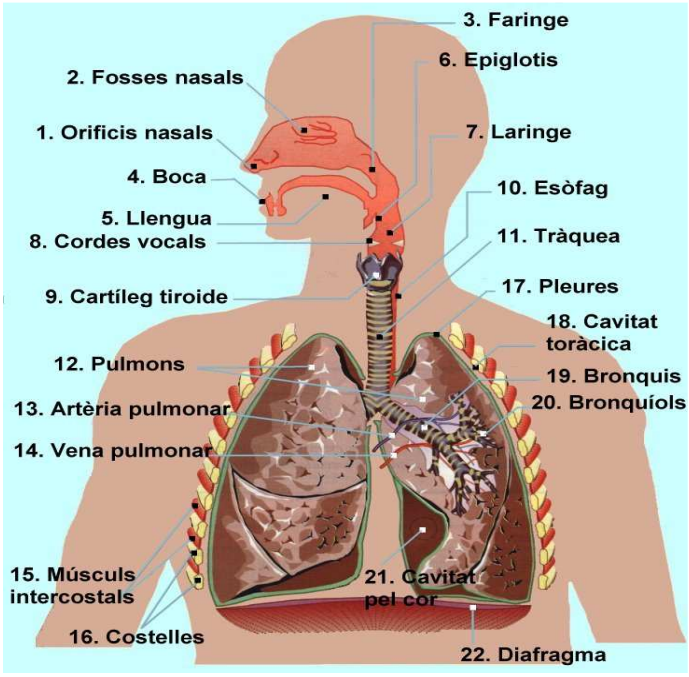


Figura 2.1 Aparell fonedor humà

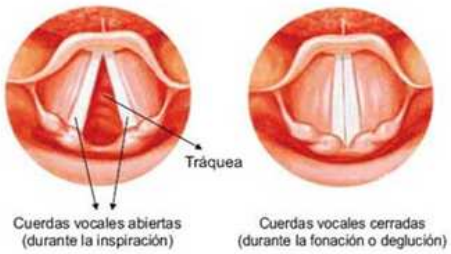


Figura 2.2 Cordes vocals

El procés de la parla comença amb la generació de l'energia suficient (flux d'aire) en els pulmons, la modificació d'aquest flux d'aire en les cordes vocals, i la posterior pertorbació per algunes constriccions i configuracions dels òrgans superiors. En el procés fonedor intervenen diferents òrgans al llarg del anomenat tracte vocal, que es troba comprés en la zona entre les cordes vocals i les obertures finals, llavis i foses nasals.

El conjunt d'òrgans que intervenen en la fonació *(veure fig. 2.1)* es poden dividir en tres grups ben delimitats:

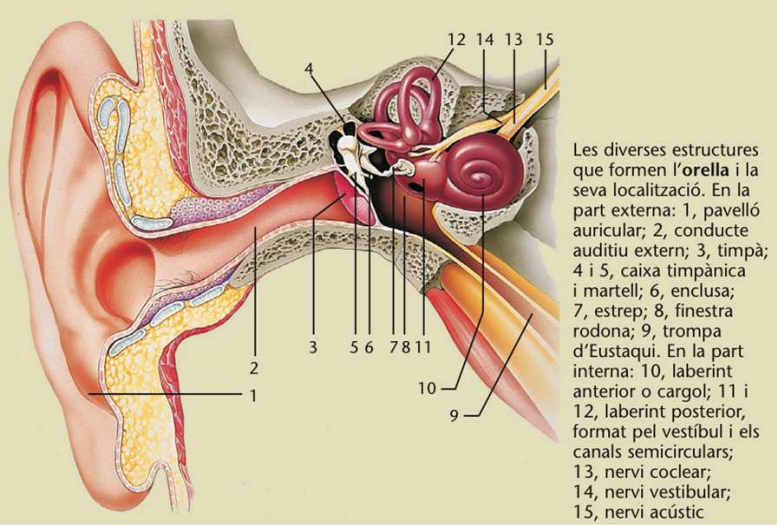
1. Òrgans de respiració (Cavitats infraglòtiques: pulmons, bronquis i tràquea).
2. Òrgans de fonació (Cavitats glòtiques:laringe, cordes vocals i ressonadors – nasal,bucal i faringi)
3. Òrgans d'articulació (cavitats supraglòtiques: paladar, llengua, dents, llavis i glotis)

Hi ha dos mecanismes bàsics de producció de veu:

1. La vibració de les cordes vocals, que dóna lloc a sons "sonors" (vocals, semivocals, nasals,etc). És un senyal modulat i quasi periòdic amb un període o freqüència fonamental anomenat *"pitch"*. Aquest senyal té una alta energia i el seu rang de freqüències està entre 300 Hz a 400 Hz.
2. Les interrupcions (totals o parcials) en el flux d'aire que surt dels pulmons, i que donen lloc als sons "sords" (fricatives, plosives, etc). El senyal d'aquest tipus es caracteritza per tenir baixa energia i component freqüencial uniforme, presentant aleatorietat com el soroll blanc.

A més també hi ha combinacions d'ambdós mecanismes, com les oclusives sonores, aquests sons es matisen després per la configuració de la resta del tracte vocal.

2.1.2 Percepció de la veu



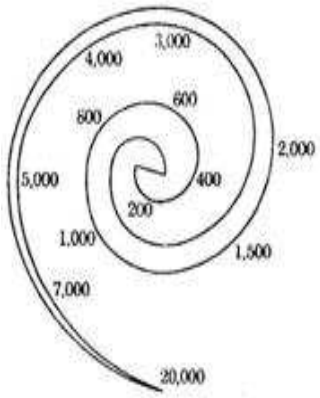


Figura 2.3 Aparell auditiu humà *Figura 2.4 Freqüències de la còclea*

Podem dividir el sistema auditiu en dos parts principals *(veure fig. 2.3)*:

- **Regió perifèrica**:

Gestiona els estímuls com a ones mecàniques. Aquesta regió està dividida en tres zones:

Oïda externa: Canalitza l'energia acústica, formada pel pavelló auricular, el conducte auditiu extern i la cara externa del timpà.

Oïda mitjana: Transforma l'energia acústica en energia mecànica, la transmet i l'amplifica a l'oïda interna.

Oïda interna: On es realitza la definitiva transformació de l'energia mecànica en impulsos elèctrics.

- **Regió central:**

És on es transformen els senyals en impulsos elèctrics i processos cognitius. Fa referència a la manera en que el cervell actua per aconseguir una òptima percepció de la veu i té els seus orígens en els impulsos nerviosos generats en l'oïda interna que conté informació sobre l'amplitud i el contingut espectral del senyal sonor.

El mecanisme físic de la percepció de la veu, s'ha constituït en un mitja de percepció molt complex i avançat. Aquest procés es realitza principalment en tres etapes: la percepció acústica dividida en captació i la transformació de l'ona mecànica, la conversió del senyal en impulsos elèctrics i l'etapa de processament neuronal, en la qual els impulsos son interpretats pel sistema nerviós.

L'orella capta les ones sonores que es transmeten a través del conducte auditiu fins al timpà. El timpà és una membrana flexible que vibra quan li arriben les ones sonores, aquesta vibració arriba a la cadena d'ossets que amplifiquen el so i ho transmet a l'orella interna a través de la finestra oval. Finalment les vibracions "mouen" els dos líquids que existeixen en la còclea *(fig. 2.4)*, deformant les cèl·lules ciliades existents en l'interior. Aquestes cèl·lules transformen les ones sonores en impulsos elèctrics que arriben al nervi auditiu i d'aquest a la corfa auditiva que és l'òrgan encarregat d'interpretar els sons.

Els éssers humans tenen una limitació en la capacitat d'audició de freqüències que va dels 20Hz als 20.000 Hz.

Les transformacions que es produeixen al senyal acústic en ser processat per l'oïda son les següents:

En l'oïda externa es produeix un filtrat passa banda de 1.5KHz a 7KHz, amb unes caigudes laterals molt suaus (10dB/octava). En l'oïda mitjana existeix un guany addicional i filtrat passa baixes amb freqüència de tall d' aproximadament 2500Hz i caiguda de 20dB/octava. En l'oïda mitjana es produeix també un control automàtic de guany (CAG): a partir de 80dB així que es disminueix el guany per sota de 2KHz. Finalment en l'oïda interna es fa un anàlisis espectral de freqüències i cada una de les cèl·lules ciliars que es correspon amb una determinada posició equival a un filtrat passa banda del senyal. Per alta freqüència la pendent dels filtres és abrupte i per baixa freqüència la pendent és suau. La cèl·lula ciliada respon de forma òptima per una freqüència determinada (ressonància), una mica menys per una freqüència baixa i molt poc per freqüències superiors a la òptima.

2.2 Fonaments del reconeixement de veu

En tot sistema de reconeixement de veu, es poden distingir quatre etapes bàsiques *(fig. 2.5)*. El mètode o tècnica utilitzada per cada una d'aquestes etapes pot variar segons l'objectiu final o pel fet que existeixen diverses formes d'obtenir els mateixos resultats o semblants. En capítols posteriors s'especifiquen les tècniques escollides en cada etapa.

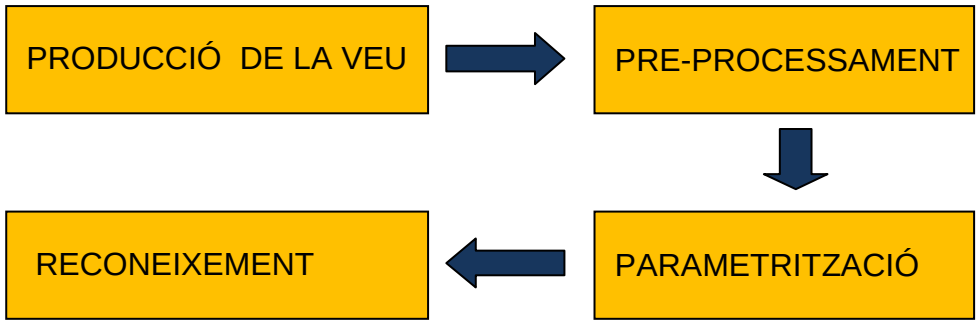


Figura 2.5 Esquema en blocs d'un reconeixedor de veu genèric

1. Producció de la veu: La veu és generada per un locutor, transformada a impulsos elèctrics per un micròfon i digitalitzada per un dispositiu convertidor Analògic / Digital.

2. Pre-processament: La fase de pre-processament del senyal de veu correspon als passos previs a la parametrització, necessaris per ressaltar les característiques més importants del senyal, així com aïllar la trama de senyal que conté informació útil i eliminar les trames de silenci o soroll.

3. Parametrització: La parametrització o extracció de característiques del senyal de veu, és una tasca important en el disseny de qualsevol sistema de reconeixement i per tant és important seleccionar la millor representació paramètrica del senyal de veu.
En aquesta etapa bàsicament es busquen dos objectius:

- Aconseguir una compressió del senyal de veu, eliminant així la informació redundant pel posterior anàlisis fonètic de les dades pre-processades.

- Realçar aquells aspectes importants del senyal que poden contribuir de manera significativa a la detecció de les diferents fonètiques.

La tècnica utilitzada aquí és l'anàlisi Cepstrum de freqüència en escala de Mel que s'explica en el punt 2.4. amb la qual es generen una sèrie de coeficients que representen les característiques del senyal de veu.

4. Reconeixement: En la fase de reconeixement s'identifiquen les paraules pronunciades a partir dels coeficients de l'etapa anterior, els quals son la base per entrenar un sistema classificador. En el nostre cas aquesta fase incorpora una etapa de comparació de patrons i una etapa de classificació mitjançant una Xarxa Neuronal.

2.2.1 Tipus de reconeixedors

Com que no existeix un mètode universal d'implementació d'un sistema reconeixement automàtic de la veu, es poden classificar segons el problema a resoldre de forma aïllada.

- **Dependents del locutor**

Els reconeixedors dependents de locutor en l'etapa d'entrenament s'utilitza el senyal de veu d'un únic locutor. És a dir s'utilitza per reconèixer únicament la veu del locutor que l'ha entrenat. Aquests sistemes són més precisos, ja que existeix menys variabilitat en els senyals a reconèixer, però presenta l'inconvenient que s'ha d'entrenar el sistema cada vegada que el vulgui utilitzar un locutor nou.

- **Independents del locutor:**

Durant l'entrenament s'utilitzen senyals o registres de molts locutors, de forma que durant el test es pot reconèixer la veu de molts locutors (sistema multi locutor), si la capacitat de generalització és bona, serà possible reconèixer paraules de locutors no

utilitzats en el procés d'entrenament, llavors estrictament estaríem parlant d'un sistema independent de locutor, que és un dels objectius d'aquest treball.

- **De paraules aïllades**

En aquest tipus de reconeixement s'han de realitzar pauses suficientment llargues entre paraules, com un dictat paraula a paraula. L'avantatge és que resulta més senzill detectar els inicis i finals de paraula que amb parla continua. Aquest és el cas del present treball, en el qual les ordres són paraules aïllades que es pronuncien amb pauses llargues.

- **De parla continua**

És el sistema més complex, ja que el reconeixedor es veurà afectat per efectes coarticulatoris: un mateix fonema no sempre sona igual; depèn de quin hagi sigut el fonema anterior. L'avantatge és que treballa amb la forma normal de parlar de les persones.

- **Segons mida del vocabulari**

Lògicament, com més gran sigui el conjunt de paraules a reconèixer, major serà el nombre d'operacions a realitzar, la memòria necessària per emmagatzemar el models i la probabilitat d'error. En el cas d'estudi es tracta d'un vocabulari petit (8 paraules).

2.2.2 Dificultats del reconeixement

El reconeixement de veu sembla tant natural i senzill per les persones que es va pensar que podria ser fàcilment realitzat per les màquines. Tanmateix, quan es va començar a profunditzar en el tema, es va comprovar que no era així. Hi ha alguns punts que compliquen el procés, el mateix fonema pronunciat per diferents interlocutors és acústicament diferent degut a variacions en la longitud del tracte vocal i la seva musculatura. Per exemple els interlocutors femenins produeixen un to més alt comparat

amb els masculins, degut a un aparell vocal més petit. A més, el mateix interlocutor pot produir versions acústicament diferents del mateix so sota diferents circumstàncies.

En particular hi ha cinc factors que determinen la complexitat del reconeixement de veu:

El Locutor : és potser l'aspecte que introdueix major variabilitat degut a que les persones no pronuncien la mateixa paraula de la mateixa forma en moments diferents. Aquesta pronunciació depèn de molts factors com l'estat d'ànim, l'entonació, el temps, la pronunciació, etc.

La forma de parlar : és el segon factor que determina la complexitat d'un reconeixedor de veu, degut a la continuïtat de la parla, és a dir una mateixa síl·laba sona diferent al principi, en el mig o al final de la paraula. D'igual forma s'ha de considerar si es realitza el reconeixement de paraules aïllades o paraules continues.

El vocabulari : és el factor que determina el nombre de paraules diferents que ha de reconèixer el sistema. Com més número de paraules més complex és el reconeixement.

La gramàtica : és el conjunt de regles que limita el número de combinacions permeses per les paraules del vocabulari.

L'entorn físic: és una part tan important com qualsevol de les anteriors per definir el reconeixedor. No és lo mateix un sistema que funciona en un ambient poc sorollós, com pot ser un despatx, comparat amb el que ha de funcionar en un avió, un cotxe o una fàbrica. També cal fer esment de l'ambient que hi ha alhora de prendre els enregistraments de veu que serviran com a patrons per l'entrenament del sistema. Com més sorollós sigui l'ambient més dificultat hi haurà alhora d'assolir taxes d'efectivitat elevades i al inrevés. Dit d'una altra manera, que si s'aconsegueix obtenir un percentatge elevat d'efectivitat, en un sistema on les mostres d'entrenament han estat presses en ambients sorollosos i diferents, significa que obtindrem un sistema més robust a la pràctica.

2.2.3 Anàlisis del senyal de veu

El senyal de veu és un senyal especial ja que mitjançant sons codifica el llenguatge parlat a part que també incorpora diverses fonts d'informació parlada com missatge, identitat, idioma, patologia, estat emocional, etc.

S'organitza jeràrquicament:

Diàleg → Frase→ Paraula→ Síl·laba → Fonema→ So

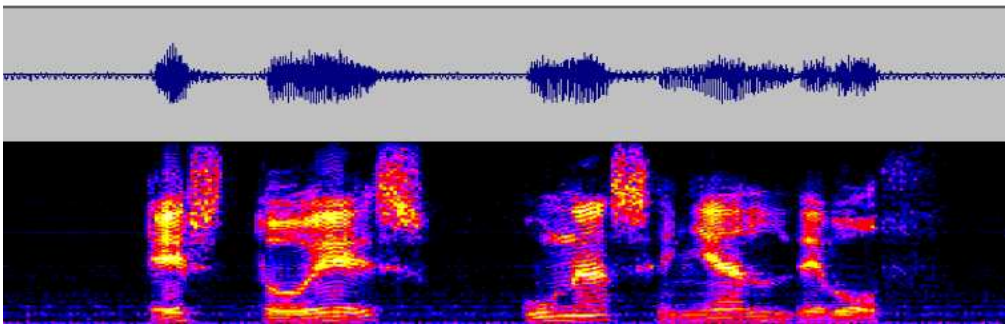


Figura 2.6 Senyal de veu i el seu espectrograma

El senyal de veu en trames llargues *(fig. 2.7)* d'uns segons de magnitud es considerada no estacionaria, és a dir les seves propietats estadístiques varien a llarg termini.

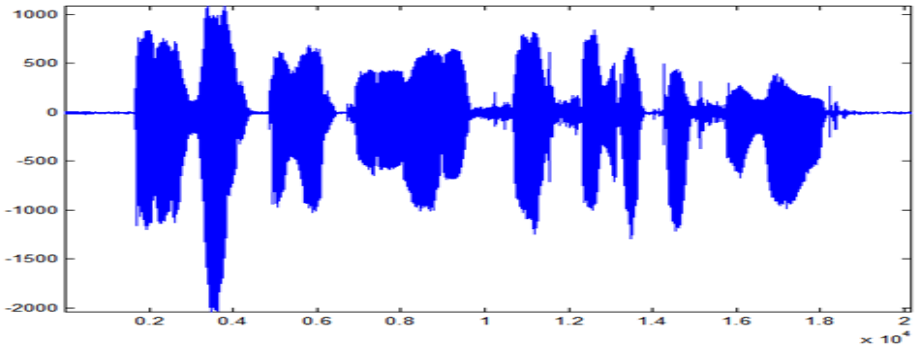


Figura 2.7 Trama llarga de senyal de veu (segons)

La no estacionarietat persisteix en trames mitges d'uns centenars de *ms.* sobretot en trams que es passa d'un soroll sonor a un de sord o en sons oclusius. En canvi en

duracions curtes (< desenes *ms.*) el senyal es comporta com quasi-estacionaria i pseudo periòdica (no generalitzable ja que poden existir trames amb aparença sorollosa).

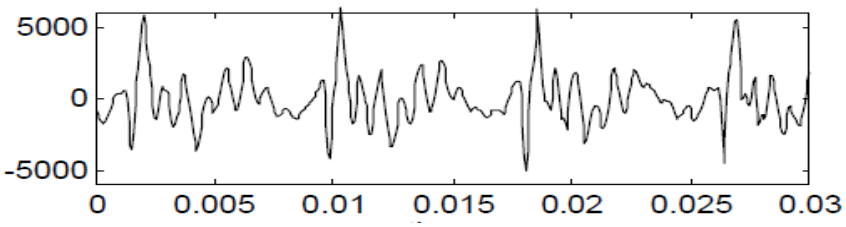


Figura 2.8 Forma d'ona d'una vocal 30ms

En el domini de la freqüència una de les característiques espectrals és la freqüència fonamental o també anomenada *"pitch"*, aquesta dona informació sobre la velocitat a la que vibren les cordes vocals en produir un so. En la *figura 2.10* podem apreciar que hi ha una diferència fonamental entre el *"pitch"* d'homes i dones, la qual cosa ens dona una indicació que a l'hora del reconeixement pot ser un factor determinant.

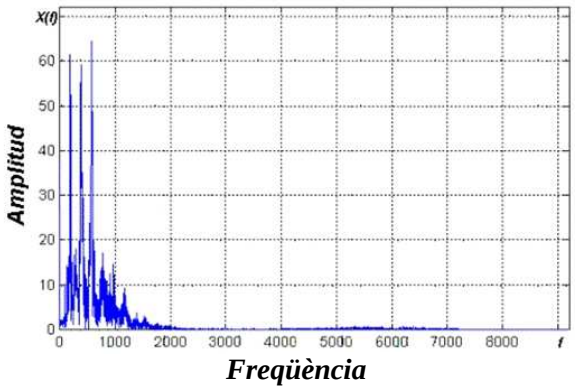


Figura 2.9 Espectre de freqüència una paraula

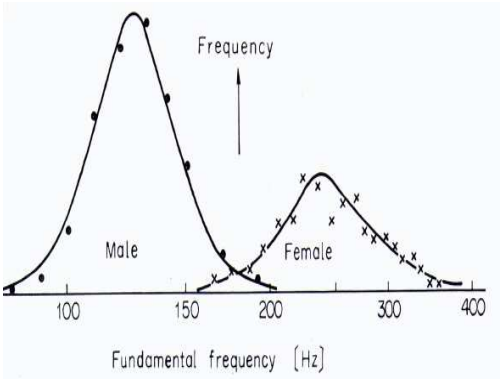


Figura 2.10 "pitch" entre homes i dones

L'espectre està conformat d'harmònics de període *"pitch"*, el qual és el rang fonamental de freqüències produïdes per les cordes vocals. Encara que l'espectre porta una gran component a la vora de la freqüència *"pitch"* (aprox. 50 Hz), té gran quantitat d'harmònics i en conseqüència té components de freqüència que s'estenen fins passats els 5 KHz.

2.2.4 Captura i digitalització del senyal de veu

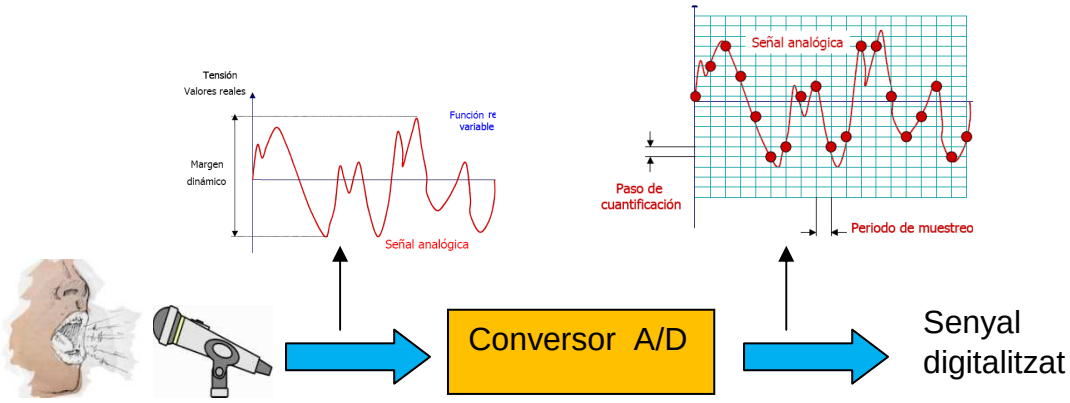


Figura 2.11 Esquema captura del senyal de veu

La naturalesa analògica de la veu introdueix la primera frontera computacional en el reconeixement de veu, aquest senyal analògic és el portador de tota la informació que es vol obtenir però al mateix temps conté una gran quantitat de soroll, composat principalment per sons que no contenen informació rellevant.

En el punt en que l'ordenador captura el senyal provinent del mon analògic i el converteix a digital *(fig. 2.11)*, es presenten les primeres pèrdues d'informació. Segons la magnitud de la freqüència de mostreig i la precisió o quantificació en determinen aquestes pèrdues.

Donat que la veu és relativament de baixes freqüències (entre 100Hz a 8Khz) i segons Nyquist sabem que és necessària una freqüència de mostreig d'almenys el doble de l'ample de banda del senyal a caracteritzar, sobre aquesta base una freqüència de mostreig de 16 KHz seria suficient. En aquest treball es va optar per una freqüència de mostreig de 44 KHz bàsicament per procurar tenir el màxim d'informació possible en les mostres de veu recollides per diversos locutors i evitar haver de repetir la operació en cas que el senyal digital d'àudio no aportés suficient informació. També per la incertesa de les necessitats d'informació d'un reconeixedor independent de locutor. A la pràctica es va fer evident que 44100Hz de mostreig generava una càrrega computacional massa elevada en l'etapa de reconeixement i es va optar per fer un delmat del senyal a 22050Hz.

Pel que fa a la quantificació més comunament utilitzada, és de 8 bits, però s'obtenen resultats molt millors amb 16 bits, que van ser els utilitzats en el treball.

2.3 Pre-processament

L'etapa de Pre-processament del senyal de veu correspon als passos previs a la parametrització o extracció de característiques, necessaris per ressaltar les seves característiques més importants i eliminar la informació innecessària com els segments de silenci del senyal de veu. En la *figura 2.12* podem veure els blocs que constitueixen el pre-processat.

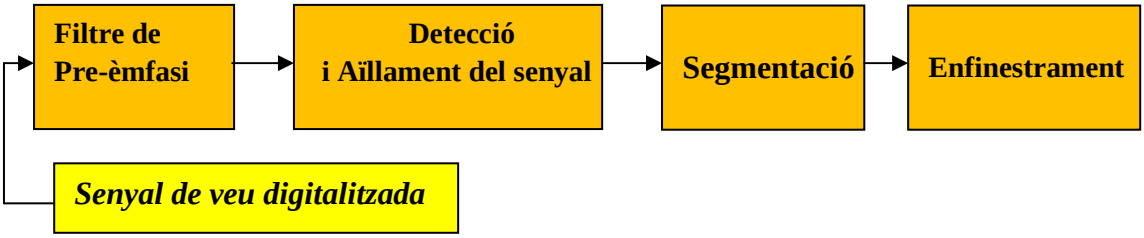


Figura 2.12 Blocs de l'etapa de pre-processament

2.3.1 Filtre de Pre-èmfasi

El pre-èmfasi s'utilitza per restaurar la pèrdua que pateix el senyal de veu en els seus components d'alta freqüència, per efecte de propagació i radiació al sortir de la cavitat vocal a través dels llavis al ambient exterior. El filtre utilitzat pot tenir coeficients fixes o ser adaptatiu, però generalment s'utilitza un filtre digital de primer ordre (FIR), la seva funció de transferència és la que es mostra en la equació.

$$H(z) = 1 - \mu . z^{-1} \qquad \textit{Equació 2.1}$$

sent $0{,}95 \leq \mu \leq 0{,}98$ (μ =0.97 en aquest treball)

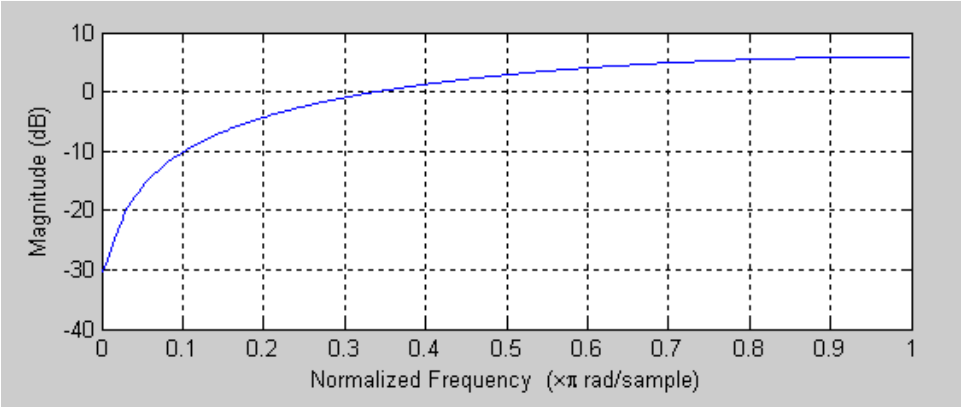


Figura 2.13 Gràfica funció transferència filtre pre-èmfasi

En Matlab es crea el filtre pre-èmfasi amb la funció *filter*:

```
b=[1 -0.97];
y=filter(b,1,x);
```

2.3.2 Detecció i aïllament del senyal de veu

En aquest apartat un cop tenim el senyal de veu digitalitzat cal realitzat un filtrat del senyal adquirit, detectant el començament i el final del senyal amb l'objectiu d'eliminar informació redundant d'entrada al sistema, és a dir eliminar silencis al començament, al final i s'escau al mig del senyal (per paraules compostes).

Es comença per fer un càlcul de l'energia del senyal d'àudio. Per això es calcula el valor absolut del senyal sencer i s'obté el pic màxim. Finalment es divideix el senyal pel valor obtingut anteriorment.

```
len = length(s); % longitud del vector
d=max(abs(s));
s=s/d;
```

El senyal normalitzat s'eleva al quadrat i es divideix pel nombre de mostres del senyal amb el que s'obté l'energia promig del senyal.

```
avg_e = sum(s.*s)/len;
```

A continuació es divideix el senyal normalitzat en finestres d'un nombre determinat de mostres, calcular l'energia d'aquest tram del senyal i si aquesta energia és major a un percentatge (llindar de decisió) de l'energia promig del senyal completa, llavors aquesta finestra o tram es conserva. En cas contrari la finestra no es considera, perquè s'interpreta que si l'energia d'aquest tram no supera el llindar establert llavors es tracta d'un interval de silenci o soroll.

```
Llindar = 0.02;
y = [0];
for i = 1:400:len-400      %Trames de 10ms aproximadament
seg = s(i:i+399);
e = sum(seg.*seg)/400 ;  % promig de cada segment o trama
if( e> Llindar*avg_e)     % si el promig energètic de cada trama es mes gran
que el
                          % promig del senyal complet pel valor llindar
   y=[y,seg(1:end)];      % es guarda en  y sinó s'elimina com espai en blanc
end
end
```

El valor del llindar correspon al 2% de l'energia promig del senyal sencer. Aquest valor es va definir després de varies proves observant el resultats d'eliminació de silenci. Les finestres es van escollir de 400 mostres que a una freqüència de mostreig de 44100Hz equival aproximadament a 10ms, així no s'eliminarà cap fonema d'àudio ja que la seva duració ronda entre els 10 i 20ms. A continuació es pot veure el senyal de veu sencer de la paraula "Amunt" *figura 2.14*, i el mateix senyal sense els silencis *figura 2.15*

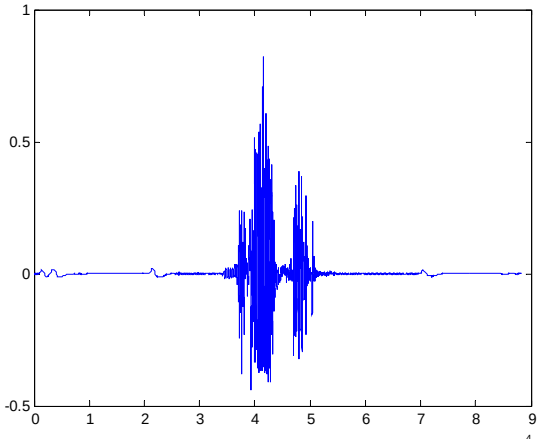


Figura 2.14 Paraula "amunt"

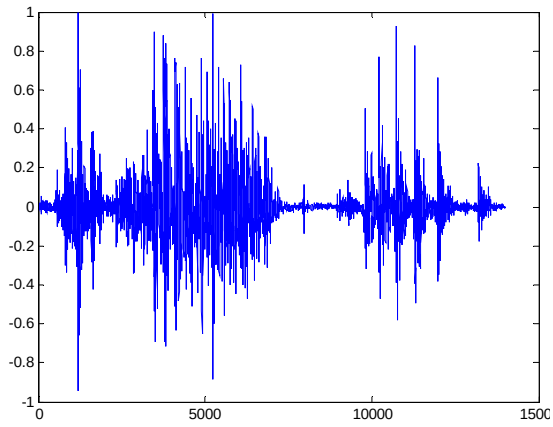


Figura 2.15 Paraula "amunt"sense silencis

2.3.3 Segmentació

L'anàlisi del senyal de veu, es desenvolupa segmentant el senyal en blocs que es tracten individualment. Normalment els segments es localitzen espaiats uniformement en el temps a intervals regulars. Es coneix que si un segment és el suficientment petit, les propietats del senyal de veu seran substancialment invariants *(veure 2.2.3)*. Això és degut a que l'aparell fonedor inverteix cert temps en la transició des d'un punt d'articulació a un altre.

Per escollir la duració dels segments és necessari considerar propietats del senyal de veu. Si s'escull una longitud de segment de entre 10 i 45 ms. hi ha forces possibilitats d'aconseguir una part representativa del senyal. Tanmateix, si l'interval de temps és massa curt, inferior a un període fonamental del senyal, es corre el risc de que les característiques interessants quedin ocultes degut a les ràpides variacions que es produeix en la part del senyal escollit.

Considerant una freqüència de mostreig de 44100 Hz que abans de la segmentació es redueix a la meitat 22050Hz (delmació), si escollíssim un valor de 512 mostres per segment, és millor potències de dos per calcular la FFT posteriorment, tindríem:
Longitud = 512 mostres / 22050Hz ≈ 23 ms (cauria dins el rang teòric)

Per evitar els efectes negatius provocats al prendre un segment que contingui una transició d'una zona del senyal quasi-estacionària a la següent, s'utilitza la tècnica del

solapament de segments. Aquest mètode consisteix en agafar una separació entre els començaments de cada segment menor que la longitud dels segments, així que es produeix un solapament entre els mateixos. Normalment la zona de solapament sol ser la meitat de la longitud del segment veure *figura 2.16*.

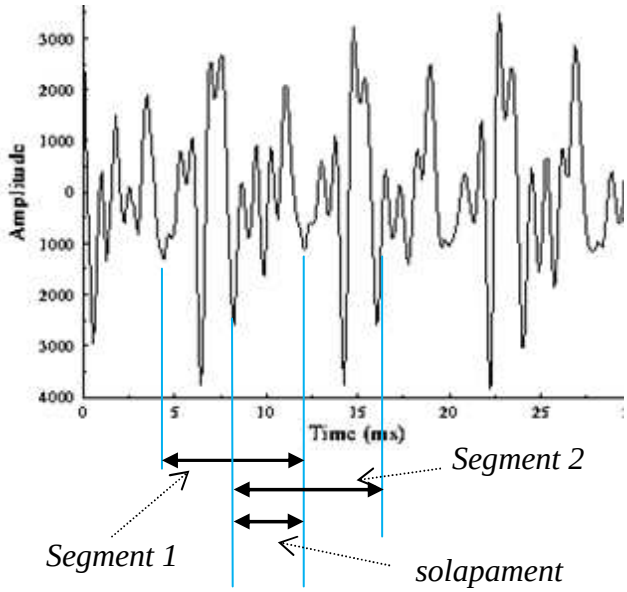


Figura 2.16 Segmentació del senyal de veu

Cal anar en compte que com més petit és el segment més llarg serà el vector de característiques MFCC de tot el senyal sencer. (veure apartat 2.4.2).

Per exemple agafant una mostra de veu de la paraula “amunt” d’un locutor a l’atzar s’han analitzat tres casos:

1. Utilitzant 512 mostres per segment i 13 coeficients *(12+Energia)* per segment, obtenim un vector de característiques MFCC de 663 coeficients pel senyal sencer.
2. En el cas de 1024 mostres per segment, obtenim un vector de característiques MFCC de 156 coeficients per tot el senyal.
3. En el cas de 2048 mostres per segment, obtenim un vector de característiques MFCC de 65 coeficients per tot el senyal.

En el nostre cas per aconseguir un comprimís entre rendiment i efectivitat es va optar per 2048 mostres per segment i 1024 mostres de solapament, per tant a una freqüència de 22050Hz (un cop delmada la freqüència original de 44100Hz) obtenim trames d'uns 92 ms. L'avantatge es que hem de processar molts menys coeficients que en el cas 1 i 2 però per contra perdem una mica de precisió o robustesa, però segons les proves fetes a la pràctica es gairebé inapreciable.

2.3.4 Enfinestrament

El procés de segmentació suposa extreure una part del senyal separant-lo de tot el conjunt, això provoca un efecte negatiu per l'anàlisi de l'evolució en el temps de les característiques del senyal. La solució a aquest problema resideix en aplicar a cada segment una funció finestra, que suavitza les vores de l'interval, fent que aquests tendeixin a zero, i ressalta la part central accentuant les propietats característiques del segment. Tan les finestres rectangulars, com les no rectangulars tenen les seves avantatges i desavantatges però en el cas específic de la finestra Hamming *(fig. 2.17)* es compleix amb el requeriment de que l'atenuació als lòbuls secundaris sigui brusca, encara que el seu lòbul principal és més ample que per una finestra rectangular, el que permet obtenir una millor resolució espectral. La finestra de Hamming es la utilitzada en aquest treball i es defineix matemàticament per l'equació:

$$W(nT) = 0{,}54 - 0{,}46 \,.\cos\left(2.\frac{n}{N}\right) \qquad 0 < n < N$$

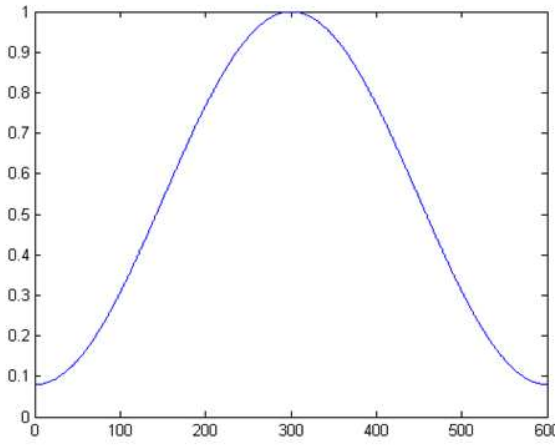


Figura 2.17 Finestra de Hamming

2.4 Extracció de característiques

L'etapa d'extracció de característiques del senyal, és bàsicament l'estimació de variables, anomenades vectors de característiques o paràmetres, a partir d'un altre conjunt de variables. Aquestes característiques obtingudes no tenen cap tipus de significat físic, i encara menys relació amb les característiques de producció de veu o també anomenades Característiques Acústiques, com per exemple el "*pitch"*, el *"jitter"*, etc.

Existeixen varies tècniques per l'extracció de característiques de veu, i poder representar paramètricament les dades del senyal de veu adquirida:

- Coeficients Cepstrals Reals (RCC), Oppenheim (1969)
- Coeficients de Predicció Lineal (LPC), Atal y Hanauer (1971)
- Coeficients Cepstrals de Predicció Lineal (LPCC), Atal (1974)
- Coeficients Cepstrals en Freqüència en escala de Mel (MFCC), Davis y Mermelstein (1980)
- Coeficients Perceptual de Predicció Linear (PLP), Hermansky(1990)

En el nostre cas utilitzarem els Coeficients MFCC *(Mel-Frequency Cepstrum Coefficients)* que estan basats en l'Espectre de Fourier, aquests son els més utilitzats en les aplicacions de reconeixement de veu, ja que tenen major robustesa al soroll i a les variacions d'estimació espectral.

2.4.1 Coeficients Cepstrals de predicció lineal (LPCC)

El mode de predicció lineal (LP), és històricament un dels mètodes més importants utilitzats per l'anàlisi de veu. La seva base fonamental és la d'establir un model filtrat per la font de d'àudio. Amb el suficient nombre de paràmetres el model de LP pot establir una bona aproximació a l'estructura espectral de qualsevol tipus de so. El mètode de LP rep aquest nom perquè pretén extrapolar el valor de la següent mostra de so x(n) com la suma ponderada de les mostres prèvies x(n-1), x(n-2), ..., x(n-k):

$$x(n) = \sum_{i=1}^{K} a_i\, x(n-1) \qquad \textit{Equació 2.2}$$

Per això s'ha de realitzar el càlcul dels coeficients, minimitzant amb alguna funció d'error *E*, concretament de mínims quadrats sobre una finestra de longitud *N*.

$$E = \sum_{n=0}^{N-1} e^2\,(n) = \sum_{n=0}^{N-1} (x(n) - \sum_{i=1}^{K} a_i\, x(n-1))^2 \quad ; 0 \le n \le N-1$$

Equació 2.3

Partint de l'Anàlisi de Predicció Lineal, és possible obtenir la expressió dels coeficients cepstrals associats (LPCC):

$c(0) = \log(1) = 0$

$$c(i) = -a(i) - \sum_{j=1}^{i-1}(1 - \frac{j}{i})\, a(j)c(i-j); \;\; 1 \le i \le N_c$$

Equació 2.4

2.4.2 Coeficients Cepstrals de Freqüència Mel (MFCC)

Una família de coeficients directament relacionada amb els LPCC son els anomenats "*mel-cepstrum*" o MFCC *(Mel Frequency Cepstral Coefficients)*, aquests van ser els utilitzats en el present treball. Son de gran utilitat en l'extracció dels paràmetres del senyal de veu, ja que estan basats en la variació coneguda dels amples de banda de les freqüències crítiques de l'oïda. Els filtres que s'apliquen al senyal en la tècnica MFCC estan espaiats linealment per freqüències menors a 1000 Hz i de forma logarítmica per freqüències majors de 1000 Hz, amb la finalitat de capturar les característiques fonètiques importants de la parla. A aquesta escala s'anomena "Escala de MEL" i la seva fórmula matemàtica és la següent:

$$Mel(f) = 2595 \cdot \log\left(1 + \frac{f}{700}\right) \qquad \textit{Equació 2.5}$$

Els passos necessaris pel càlcul dels MFCC es mostren en el següent diagrama *(fig. 2.18)*:

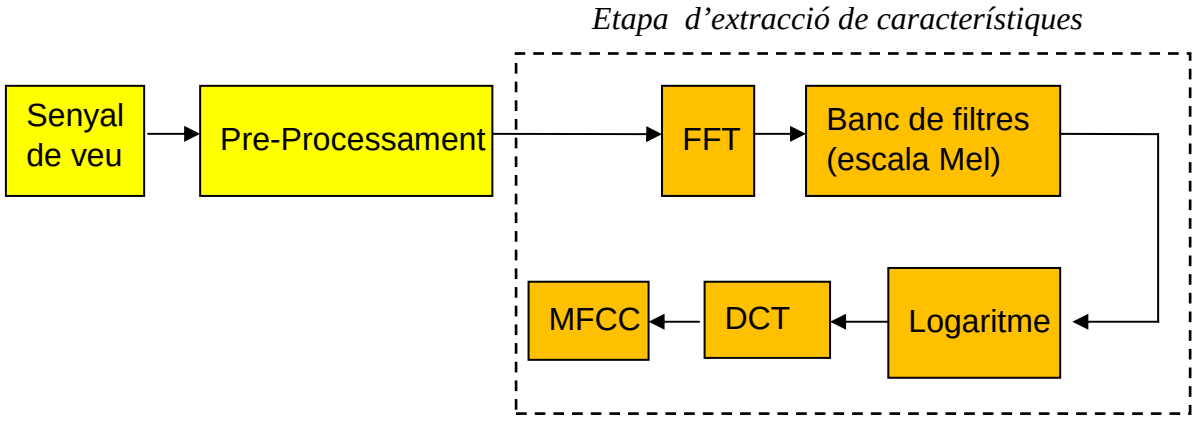


Figura 2.18 Pre-processament i extracció dels coeficients MFCC

- El primer bloc calcula la transformada de Fourier de temps curt a cada una de les trames obtingudes en l'etapa de pre-processament mitjançant l'equació:

$$X(n,\omega_k)=\sum_{m=-\infty}^{\infty} x(m)\cdot w(n-m)\cdot e^{-j\omega_k m} \qquad \omega_k=\frac{2\pi}{N}\cdot k$$

Equació 2.6

El quadrat de la magnitud de $X(n,\omega_k)$ és ponderat per una serie de filtres distribuïts sobre l'escala de Mel per després calcular la anomenada "log-energia" del filtre l-èssim mitjançant la eqüació:

$$E_{Mel}(n,l)=\frac{1}{A_l}\cdot\sum_{k=L_l}^{U_l}\left|V_L(\omega_K)\cdot X(n,\omega_K)\right|^2$$

Equació 2.7

On L_1 i U_1 son les freqüències de tall inferior i superior del filtre l-èssim

- El segon bloc, el banc de filtres, és el encarregat de modelar la resposta auditiva humana espaiant les bandes de freqüències de manera logarítmica (escala de Mel) de la mateixa manera que ho fa la còclea del nostre sistema auditiu humà. El banc

de filtres en escala de Mel té la forma que s'indica en la *figura 2.19* i els filtres que el conformen poden ser triangulars o tenir altres formes, tals com Hamming, Hanning, Kaizer, etc. El triangular és el més utilitzat però en la pràctica el que millors resultats va donar en aquest treball és el de Hamming.

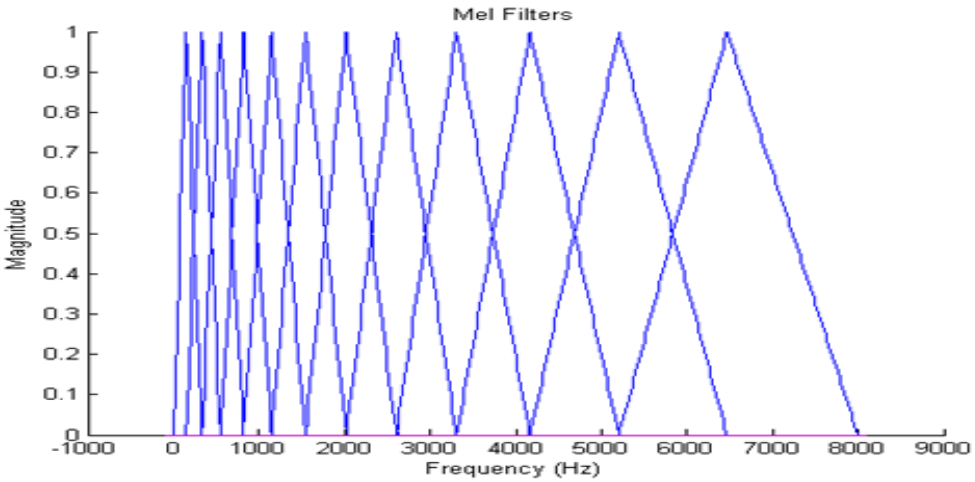


Figura 2.19 Banc de filtres espaiats linealment en escala Mel

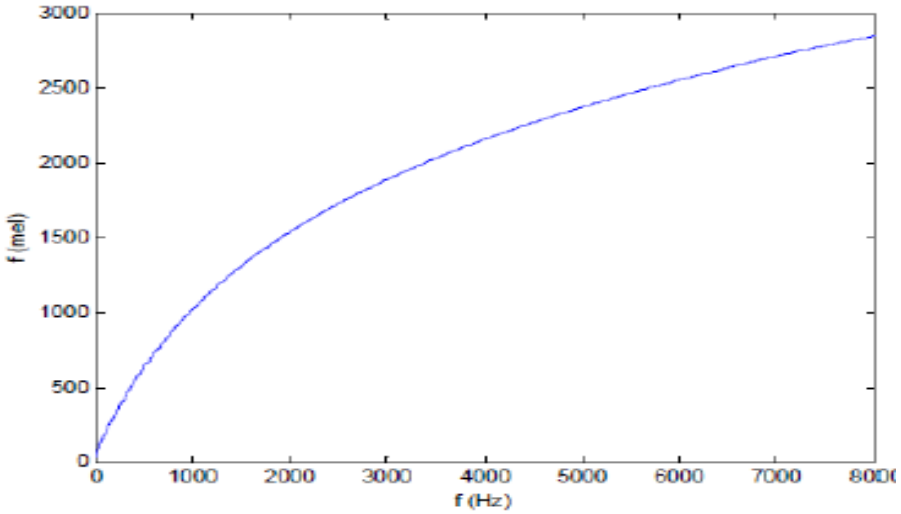


Figura 2.20 Representació de l'escala de Mel

- Finalment es converteix l'espectre logarítmic Mel novament al domini del temps utilitzant la Transformada Discreta del Cosinus, donat que els coeficients cepstrals son números reals. El càlcul d'aquests coeficients es realitza mitjançant l'equació.

$$C_{Mel}[n,m] = \frac{1}{R} \cdot \sum_{l=1}^{R} \log\{E_{Mel}(n,l)\} \cdot \cos\left[n\left(l - \frac{1}{2}\right) \cdot \frac{\pi}{l} \right], \qquad \text{n=1,2,3…K}$$

Equació 2.8

On *K* és el número de coeficients cepstrals, que en general se sol escollir entre 10 i 20, en el nostre cas es van escollir 12 coeficients. A aquests coeficients s'hi afegeixen les seves derivades en el temps per considerar els canvis temporals en l'espectre del senyal i l'energia de la trama com: $logE = \sum_{n=1}^{N} x(n)^2$. Aquestes derivades s'anomenen coeficients delta (primera derivada) i acceleració o doble delta (segona derivada). La raó perquè s'utilitzen també aquest paràmetres, es basa en el fet que en sistemes independents de locutor com es el nostre cas, les freqüències de ressonància (formants) fluctuen considerablement d'uns locutors a altres, mentre que les variacions d'aquestes freqüències (pendents de les formants) son més semblants. Així doncs els coeficients delta es calculen com segueix:

$$\Delta c(n) = \frac{1}{\sum_{i=1}^{D} i^2} \sum_{i=1}^{D} i * (c(n+i) - c(n-i))$$

Equació 2.9

On c(n) son els coeficients MFCC per cada trama, D normalment té el valor de 2. Els coeficietns doble delta $\Delta\Delta c(n)$ es calculen de forma similar però a partir dels delta. Així doncs per exemple i en el nostre cas, tindríem un vector de característiques de 39 coeficients format com segueix:

12 coeficients MFCC + Energia + 12 coeficients delta + 1 energia delta + 12 coeficients doble delta + 1 energia doble delta = ***39 coeficients*** *en total per cada trama del senyal.*

Els coeficients delta i doble delta son importants per millorar la precisió global d'un sistema de reconeixement de veu, però aquest enfocament incrementa la dimensió del vector de característiques i per tant fa augmentar la carga computacional en l'etapa

de reconeixement. Una solució a aquest problema és realitzar una suma ponderada de coeficients anomenada *WMFCC (weighted MFCC)* com segueix:

$$wc(n) = c(n) + p * \Delta c(n) + q * \Delta\Delta c(n) \qquad \textit{Equació 2.10}$$

Com que aquestes característiques delta i doble delta contribueixen en menor mesura que els coeficients MFCC, son ponderades amb valors de p i q restringits a $q<p<1$. Així doncs en el nostre cas utilitzant 12 coeficients MFCC + Energia, amb q=1/6 i q=1/3 tenim un vector de dimensió 13 de característiques en lloc de 39

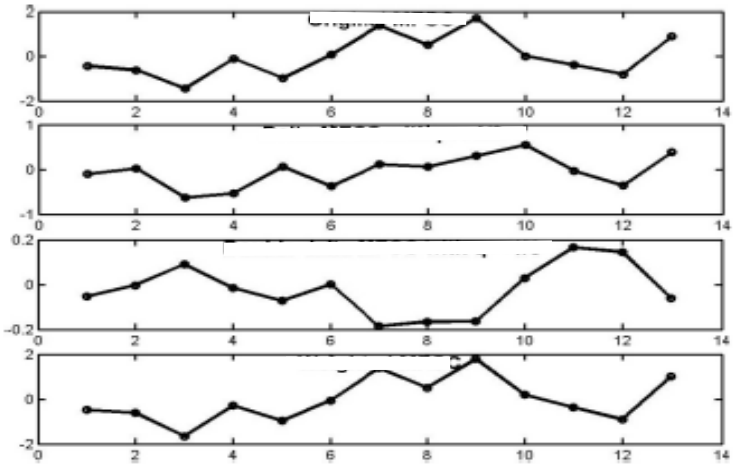


Figura 2.21 MFCC ponderats (dimensió 13) d'una trama del senyal de veu

Mitjançant el procés descrit, a una freqüència de mostreig de 22050 Hz, per cada trama de veu de duració aproximada igual a 92 ms. amb solapament, es calcula el conjunt de coeficients cepstrals. Aquest és el resultat de la Transformada Discreta de Cosinus de la Densitat Espectral de Potencia expressada en l'escala de Mel. A aquest conjunt de coeficients s'anomena *Vector Acústic o Vector de característiques*, que representa les característiques més importants de la veu pel reconeixement.

2.5 Tècniques de reconeixement

Per a desenvolupar un sistema de reconeixement per a paraules aïllades d'un vocabulari restringit i de poques paraules, la millor opció és la tècnica de comparació de patrons. La principal avantatge resideix en que no és necessari descobrir característiques espectrals de la veu a nivell fonètic, per tant evita desenvolupar etapes complexes de detecció de formants, trets distintius dels sons, to de veu, etc.

Amb aquesta tècnica serà necessària la generació d'uns models o patrons de referència, un bloc de comparació de patrons i un bloc de classificació o lògica de decisió *(veure fig. 2.22)*. Per dur a terme aquests processos es poden escollir diverses tècniques, però les més utilitzades son: els Models de Mescles de Gaussianes (GMM) *(veure 2.5.1),* l'algoritme d'Alineament Temporal Dinàmic (DTW) *(veure 2.5.2),* les Xarxes Neuronals (ANN) *(veure 2.5.3),* la Quantització Vectorial (VQ) i els Models Ocults de Markov (HMM) o bé combinacions entre elles. Pel treball dut a terme, s'han fet proves amb GMM i ANN, per separat i conjuntament, però al no donar resultats satisfactoris es va optar finalment per implementar un reconeixedor amb molt bons resultats combinant l'Alineament Temporal Dinàmic per crear els models de referència i una Xarxa Neuronal Artificial com a classificador *(veure 3.3.4)*

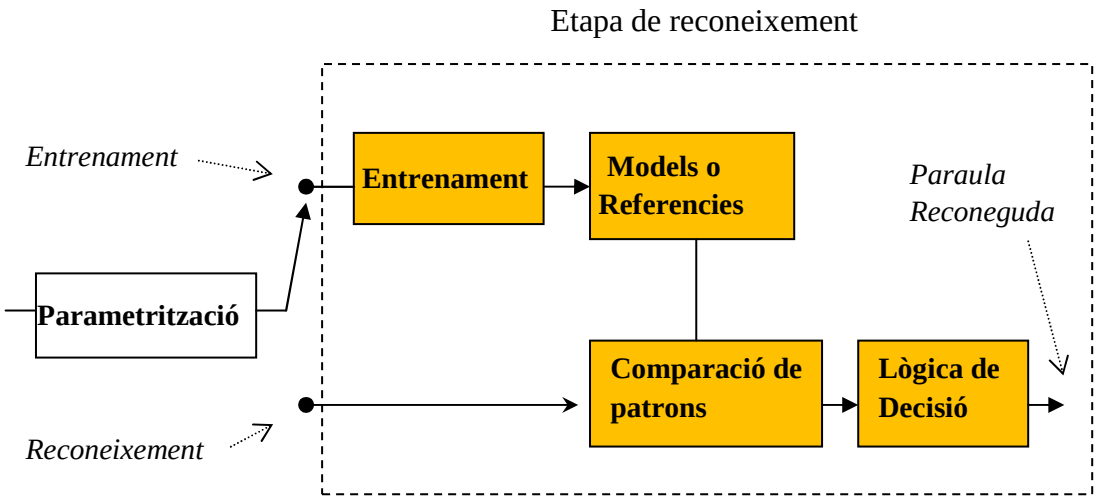


Figura 2.22 Reconeixement de veu basat en patrons

2.5.1 Models de Mescles de Gaussianes (GMM)

Els Models de Mescles de Gaussianes GMM *(Gaussian Mixture Models)* són àmpliament utilitzats pel modelat de la veu a partir dels vectors de característiques (en el nostre cas MFCC adquirits de cada classe o paraula. Una vegada obtinguda certa quantitat d'aquests vectors per cada paraula, es crea un model probabilístic que el representi de forma singular.

Donat un vector de característiques: x, la mescla de densitats Gaussianes s'expressa com:

$$P(x|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \boldsymbol{p_i b_i(x)} \qquad \textit{Equació 2.11}$$

Que no és més que la combinació lineal ponderada de M densitats Gaussianes b_i i que representa la probabilitat d'observar un determinat vector de característiques x de certa paraula $\boldsymbol{\lambda}$, on:

- x és el vector de dimensió D a observar
- w_i son els pesos de cada component Gaussiana i compleixen $\sum_{i=1}^{M} w_i = 1$
- $b_i(x)$ son les densitats Gaussianes D-dimensionals, cada una amb la forma:

$$b_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i),\right\} \quad \text{amb i: 1…M}$$

Equació 2.12

μ_i i $\sum_i$ son el vector de mitjanes i la matriu de covariància corresponen a la i-èssima mescla

- M és l'ordre del model o nombre de Gaussianes que té el model

D'aquesta forma cada paraula estarà representada per un model de mostres Gaussianes $\boldsymbol{\lambda}$ els seus paràmetres son: $\{w_i,\ \mu_i,\ \sum_i\}$ amb $i = 1…M$

Alhora d'escollir la quantitat de M mescles Gaussianes amb la qual es modelarà, si s'escull un nombre elevat pot provocar que el model trobat sobre ajusti excessivament a les dades. Per altre banda si s'escull M petit pot portar a que el model no sigui prou diferent als altres i llavors no es pugui reconèixer la paraula.

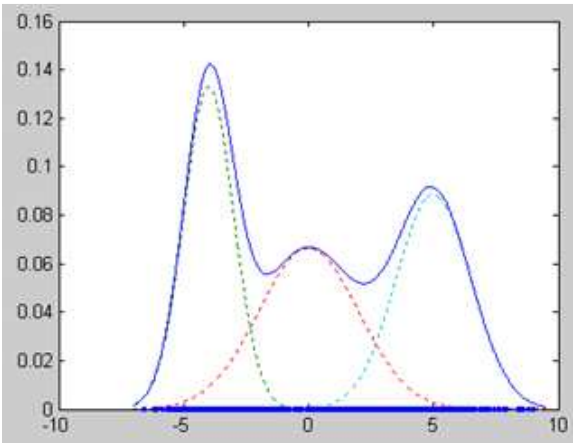


Figura 2.23 Model GMM de 3 Mescles Gaussianes

Procés d'entrenament:

A partir d'una col·lecció de vectors X = $\{x_1 \dots x_T\}$ d'entrenament d'una paraula, s'estimen els paràmetres del model utilitzant l'algoritme EM (estimació – maximització).

Partint d'un model inicial, l'algoritme EM refina iterativament el model GMM incrementant la seva versemblança, fins que arriba a un nivell de convergència predeterminat.

En general el conjunt de vectors de característiques és molt gran, i per tant, els vectors de P(...) son sovint molt petits. Per aquesta raó és comú calcular el logaritme de la versemblança que ve donat per:

$$Log\ P(X|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log P(x_t|\lambda) \qquad \textit{Equació 2.13}$$

A aquest valor s'anomena Logl (Log-Likelihood) i és la mesura que ens diu la probabilitat de que els vectors X pertanyin al model $\boldsymbol{\lambda}$

La condició per parar la iteració pot ser : $LogP\left(X|\lambda^{(k)}\right) - \log P\left(X|\lambda^{(k-1)}\right) < \varepsilon$ o bé es pot imposar un número màxim d'iteracions.

Procés de Test

El *Logl* s'utilitza per mesurar el nivell d'ajust d'un model a les dades experimentals. Llavors el sistema suposa que els vectors x d'entrada pertanyen a la paraula que ja té el seu model corresponent λ creat en la base de dades. Només cal avaluar el *Logl* per cada model i aquell amb el *Logl* més elevat és el que té més probabilitats de correspondre a aquell model.

2.5.2 Alineament Temporal Dinàmic (DTW)

El problema que es presenta quan es pronuncia una paraula és que aquesta no sempre es fa a la mateixa velocitat, el que produeix importants distorsions temporals. Aquestes distorsions afecten no només a la paraula considerada sinó també als seus components acústics. Les variacions temporals no son generalment proporcionals a la velocitat de locució i podran variar de locutor a locutor. Un algoritme que permeti realitzar la alineació entre paraules diferents és el DTW (Dynamic Time Warping). Aquesta tècnica s'encarrega de realitzar la comparació de patrons acústics, tenint en compte la variació en l'escala del temps de dues paraules a comparar.

Per aplicar la tècnica de DTW s'escull una representació o varies representacions de cada paraula que serviran de patrons o referències. Les referències son els coeficients MFCC extrets en l'etapa de parametrització que es guarden per ser comparats amb els coeficients MFCC de la paraula a reconèixer.

Per poder comparar les dues paraules (paraula de referència coneguda amb paraula desconeguda) el DTW utilitza la distància euclidiana que mesura la distorsió entre elles. És a dir per cada paraula desconeguda s'obtenen totes les distàncies euclidianes entre aquesta i totes les referències prototipus, així que es pot fer un simple classificador escollint com a paraula reconeguda la paraula de referència que té la mínima distancia amb la paraula a reconèixer.

Les cadenes de la paraula (desconeguda) i la paraula coneguda (referència) es col·loquen sobre els eixos I i J respectivament, de manera que els primers components de cada vector queden en el primer punt de cada eix

Vectors paraula referència = { i(1),..... i(I)}
Vectors paraula desconeguda= {j(1).....j(J)}

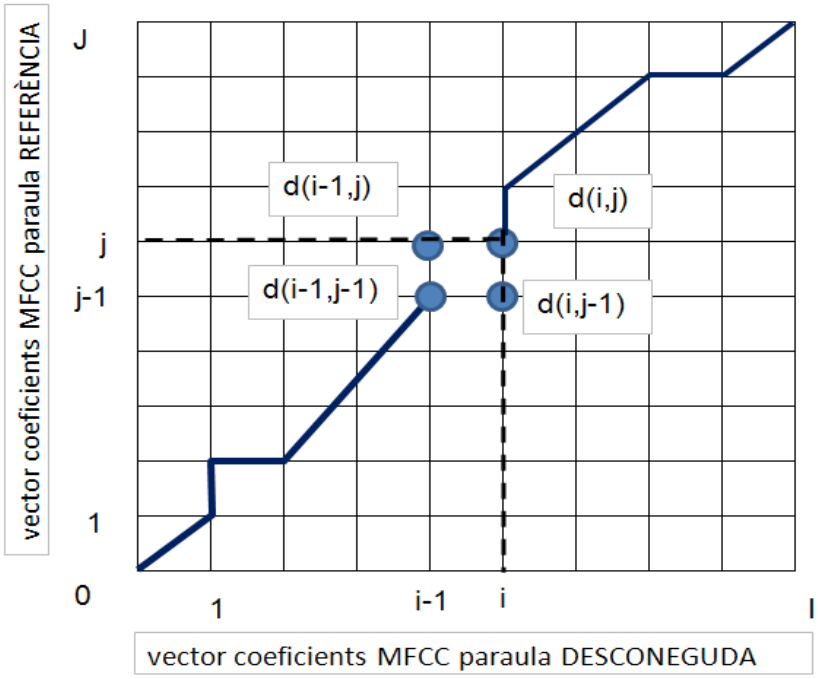


Figura 2.24 Càlcul distancies DTW

Cada node en el pla és un número real positiu. El problema consisteix en trobar el camí que recorri el pla amb la distancia més curta. Aquest camí ha de començar el punt origen (0,0) i ha d'acabar en el node final (I,J). Les distancies o costos son assignats als camins prenen com a base els punts anteriors. Per exemple en la figura per arribar al punt d(i,j) es pot agafar d(i-1,j), d(i-1,j-1) o d(i,j-1).

Si definim els valors d(i,j) corresponents a les distancies entre els vectors i-èssim de la paraula desconeguda, composta per I vectors i j-èssim de la paraula de referència composta per J vectors, aquestes formen una matriu com la figura 2.24. L'alineament de les dues seqüències de vectors es correspon amb un camí en la matriu de distancies (indicat amb traç gruixut) que parteix de l'element d(1,1) i finalitza d(I,J). La distancia

total entre els vectors alineats correspon a la suma dels elements de la matriu de distancies obtingudes en el camí, de forma que la cerca de l'alineament òptim és equivalent al camí amb menor distancia total.

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Equació 2.14 Distància euclidiana

2.5.3 Xarxes Neuronals Artificials (XNA)

Les Xarxes Neuronals Artificials tracten de modelitzar esquemàticament l'estructura hardware del cervell, per reproduir les seves característiques computacionals. Aquests sistemes de processament d'informació paral·lels, distribuïts i adaptatius, son capaços d'aprendre de l'experiència a partir de dades de l'entorn i utilitzant algoritmes numèrics.

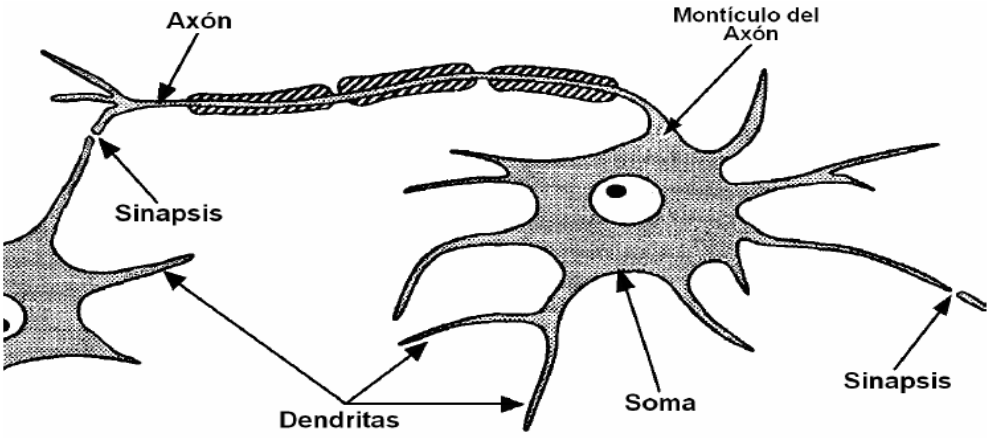


Figura 2.25 Principals components d'una neurona

Una neurona biològica és una cèl·lula especialitzada en processar informació. Està composta pel cos de la cèl·lula (soma) i dos tipus de ramificacions: l'axó i les dendrites. La neurona rep senyals (impulsos) d'altres neurones a través de les seves dendrites i transmet senyals generades pel cos de la cèl·lula a través de l'axó.

Les neurones es comuniquen entre sí mitjançant trens de polsos de curta durada (milisegons). El missatge, està modulat en la freqüència de transmissió dels polsos. Aquesta freqüència varia sobre els 100Hz. Per tant és més d'un milió de vegades inferior a la velocitat de commutació dels circuits electrònics típics d'avui en dia.

Tanmateix, l'ésser humà és capaç de realitzar tasques complexes en un temps inferior i amb un percentatge d'encerts superior al aconseguit actualment amb ordenadors. Per exemple el cas que ens ocupa, en tasques de reconeixement de veu, o de locutor a través de la veu, és molt superior a qualsevol màquina. Això és degut a la gran capacitat de treballar en paral·lel i que està extremadament lluny de poder ser implementada en un circuit electrònic.

Les XNA (Xarxes Neuronals Artificials) son sistemes de processament d'informació que la seva estructura i funcionament estan inspirades en les xarxes neuronals biològiques. En tot model de XNA es troben quatre elements bàsics.

- Un conjunt de connexions, pesos o sinapsis que determinen el comportament de la neurona, aquest valors es multipliquen amb la seva respectiva entrada les quals poden ser excitadores si presenten un signe positiu (connexions positives) i les inhibidores (amb signe negatiu).
- Una funció que s'encarrega de sumar totes les entrades multiplicades pels seus pesos corresponents, formant un senyal resultant *n*.

$$n = \sum_{i=1}^{N} W_i X_i \qquad \textit{Equació 2.15}$$

- Una funció d'activació que pot ser lineal o no lineal utilitzada per limitar l'amplitud de la sortida de la neurona.

$$y = f\left(\sum_{i=1}^{N} W_i X_i + b\right) \qquad \textit{Equació 2.16}$$

- Un guany exterior que determina el llindar d'activació de la neurona.

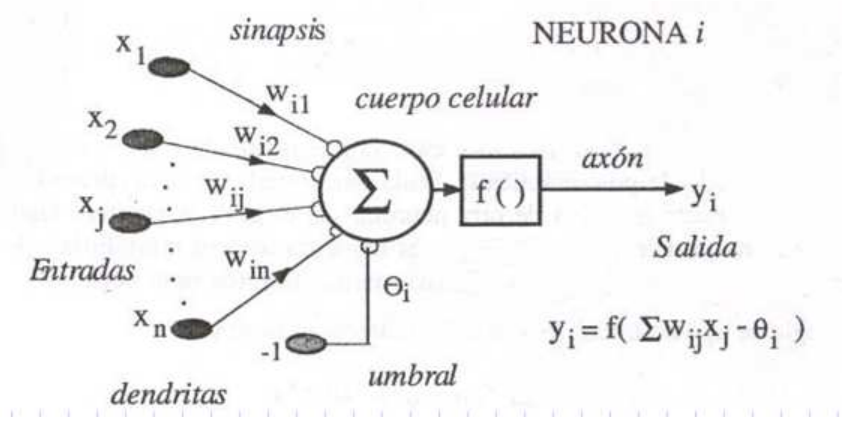


Figura 2.26 Model estàndard d'una neurona artificial

En la figura 2.26 Podem veure les funcions d'activació més utilitzades en xarxes neuronals.

	Función	Rango	Gráfica
Identidad	$y = x$	$[-\infty, +\infty]$	
Escalón	$y = sign(x)$ $y = H(x)$	$\{-1, +1\}$ $\{0, +1\}$	
Lineal a tramos	$y = \begin{cases} -1, & si\ x < -l \\ x, & si\ +l \le x \le -l \\ +1, & si\ x > +l \end{cases}$	$[-1, +1]$	
Sigmoidea	$y = \frac{1}{1+e^{-x}}$ $y = tgh(x)$	$[0, +1]$ $[-1, +1]$	
Gaussiana	$y = Ae^{-Bx^2}$	$[0,+1]$	
Sinusoidal	$y = A\,\mathrm{sen}(\omega x + \varphi)$	$[-1,+1]$	

Figura 2.27 Funcions d'activació de les xarxes neuronals

Una xarxa es pot connectar amb qualsevol tipus de topologia. Cada tipus de topologia s'utilitza per algun tipus particular d'aplicació

- **Xarxa Auto-asociativa**: És la més utilitzada pels problemes de completar patrons.
- **Xarxa en Capes**: És molt utilitzada per l'associació i reconeixement de patrons.
- **Xarxes Recurrents**: Son bastant útils per realitzar seqüències de patrons.
- **Xarxes Modulars**: Serveixen per la construcció de sistemes complexos a partir de components simples.

Pel reconeixement de patrons com és el nostre cas en el que cal reconèixer paraules s'utilitzarà una xarxa en capes *(veure fig. 2.28*)

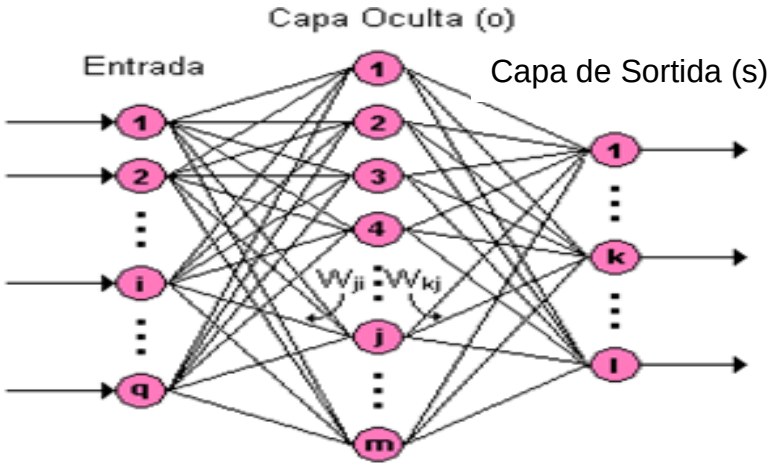


Figura 2.28 Xarxa Neuronal de tres capes

L'aprenentatge és el procés pel qual una xarxa neuronal modifica els seus pesos en resposta a una informació d'entrada. Durant aquest procés els pesos de les connexions de la xarxa tenen modificacions, afirmant que el procés ha finalitzat quan els valors dels pesos es mantenen estables.

Existeixen dos tipus d'aprenentatge: el supervisat i el no supervisat.

L'aprenentatge supervisat es caracteritza perquè els procés d'aprenentatge es realitza mitjançant un metòdic entrenament controlat per un agent extern, conegut com supervisor, que determina la resposta que hauria de generar la xarxa a partir d'una entrada determinada. El supervisor comprova la sortida de la xarxa i si aquesta no coincideix amb el que es vol, es procedeix a modificar els pesos de les connexions per aconseguir que la sortida s'aproximi a la desitjada.

3. MARC EXPERIMENTAL

3. MARC EXPERIMENTAL

3.1 Recursos utilitzats

Matlab: (abreviatura de Matrix Laboratory) és un software matemàtic especialment útil per a la manipulació de matrius, la representació de dades, funcions, implementació d'algoritmes, creació d'interfícies d'usuari (GUI) i la comunicació amb programes en altres llenguatges. Disposa d'un entorn de desenvolupament integrat (IDE) amb un llenguatge de programació propi (llenguatge M).

Voicebox: (http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html) és un *Toolbox* de processament de veu amb rutines en Matlab especialment dedicat al anàlisis i síntesis de senyals de veu. Aquest *toolbox* incorpora la funció "melcepst" amb la qual s'han generat els coeficients MFCC dels senyals de veu, entre d'altres rutines.

Micròfon: Un micròfon de gama mitja-baixa amb el qual s'han enregistrat totes les mostres de veu per confeccionar la base de dades o corpus de veus.

Tarja de so: Una tarja de so estàndard inclosa en un portàtil per digitalitzar el senyal de veu

Algoritme DTW: Un script en Matlab que executa l'Alineament Temporal Dinàmic entre dos vectors de característiques. Aquest algoritme s'ha obtingut de la pàgina Matlab Central:
http://www.mathworks.com/matlabcentral/fileexchange/6516-dynamic-time-warping

3.2 Elaboració del corpus de veus

El corpus de veus és el conjunt d'arxius amb informació de veu digitalitzada, del qual s'extreuen el conjunt de paràmetres per entrenar i testejar el sistema desenvolupat. Els fitxers d'àudio es van capturar en format *wav* , de la següent manera:

- La Freqüència de mostreig de la captura d'àudio va ser de 44100Hz i 16 bits de quantificació
- Es van escollir 8 paraules : *"Dreta", "Amunt","Esquerra","Avall", "Giradreta", "Giraesquerra", "Engega", "Para"*
- Cada persona va pronunciar cinc repeticions per cada paraula a reconèixer.
- Es van enregistrar un total de 32 persones o locutors, 14 homes i 18 dones amb edats compreses entre els 24 i els 65 anys.
- En total es van obtenir 32 locutors x 5 registres x 8 paraules = 1.280 mostres.

És important, durant l'etapa de gravació de sons, tenir en compte les condicions de soroll del lloc, distancia del micròfon, ajust del guany i sensibilitat del micròfon, per obtenir taxes de reconeixement elevades.

En el cas del present treball no es van tenir en compte aquestes condicions ja que les mostres van ser preses en llocs físics diferents, amb sorolls de fons diferents, el micròfon no era de gama alta i tampoc es va prestar especial atenció a la distancia del micròfon al locutor ni al guany. Així el sistema es va desenvolupar tenint en compte les condicions reals desfavorables amb la que normalment han de treballar els sistemes de reconeixement, les quals repercuteixen en la qualitat de les dades d'àudio. Tot i així com es veurà en l'apartat 3.5, els resultats son molt satisfactoris, els quals encara haguessin millorat molt més amb condicions de gravació òptimes.

3.3 Disseny i implementació de l'etapa de reconeixement

En el procés d'implementació de l'etapa de reconeixement es van provar diverses tècniques per obtenir uns models o referències de cada paraula i també diverses formes de classificació. En els següents apartats es detalla cada tècnica utilitzada.

3.3.1 Reconeixedor basat en Models de Mescles Gaussianes

En el desenvolupament del reconeixedor basat en GMM, per generar els models de cada paraula s'ha utilitzat 11 Locutors per crear els Models (5 homes i 6 dones), cada locutor va registrar 5 vegades cada paraula i per tant s'han obtingut 55 enregistraments per cada paraula. En la fase de Test es van utilitzar 10 locutors independents (5 dones i 5 homes) dels locutors utilitzats per crear els models.

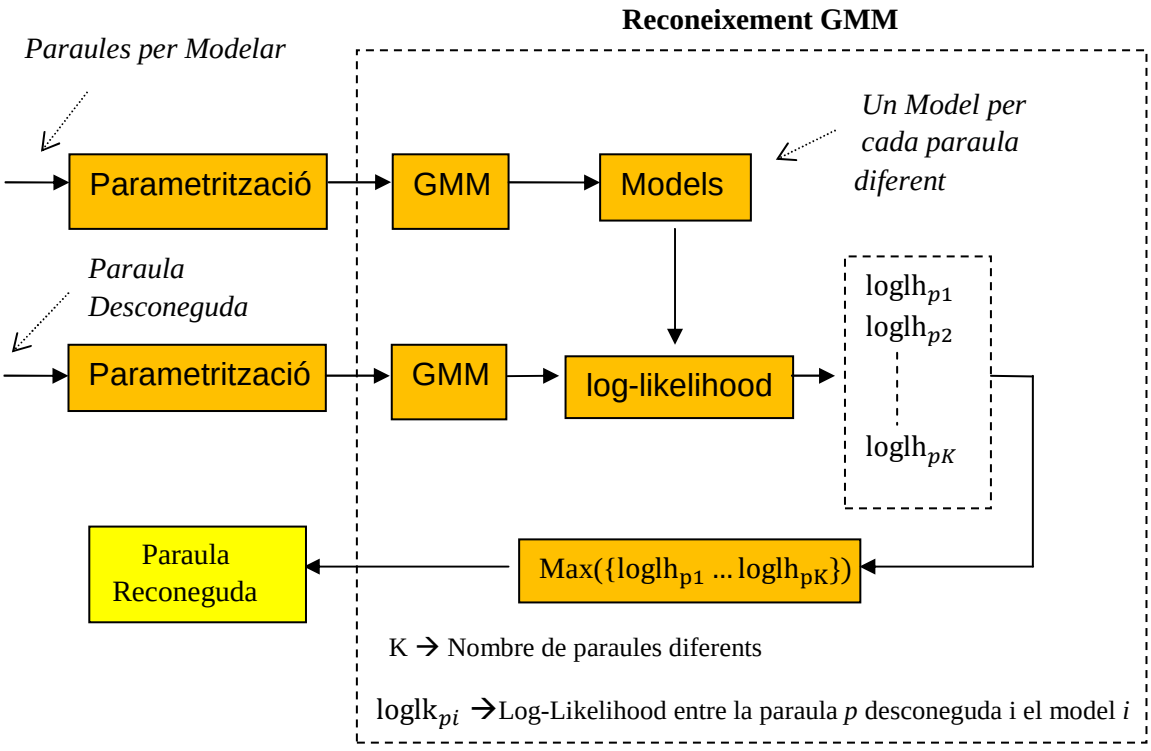


Figura 3.1 Esquema Reconeixedor de veu amb GMM

1. Abans de la parametrització del senyal de veu aquest passa per l'etapa de pre-processament tal com s'ha detallat en l'apartat 2.3

2. En l'etapa de parametrització s'obtenen les matrius de coeficients MFCC de cada enregistrament d'una paraula. Després es concatenen aquestes matrius de coeficients en una sola matriu per generalitzar al màxim el model. El procés es repeteix per cada paraula del vocabulari permès, finalment s'obtenen tantes matrius com paraules a modelar.

3. En l'etapa de Reconeixement s'obtenen primer de tots els Models a partir de les matrius de coeficients de l'apartat anterior, per fer-ho s'utilitza la funció de Matlab :

```
Models{k}=gmdistribution.fit(CoefTotal(:),M,'CovType'
,'diagonal','Options',options,'Regularize',0.01);
```

La funció gmdistribution.fit obté un model de mescles de Gaussianes utilitzant l'agoritme EM (Expectation-Maximization).

- ***CoefTotal(:)*** *:* és la matriu de coeficients MFCC total obtinguda en el pas 2.
- **M** : és el nombre de Gaussianes utilitzades per ajustar el model a les dades, que en el nostre cas els millors resultats es van donar per M=6.
- ***'CovType', 'diagonal'***: indica que es restringiran les matrius de covariança a ser diagonals, això redueix el nombre de paràmetres obtinguts.
- *Models{k}:* és una matriu de *cell array* on es guarden els paràmetres del model obtingut, aquest paràmetres son per cada Gaussiana:

  Sigma ($\boldsymbol{\sigma^2}$): Covariança

  mu ($\boldsymbol{\mu}$): Mitja

  PComponents: Proporcions de mescla de cada Gaussiana

4. Un cop obtinguts els Models s'apliquen els mateixos passos del 1 al 3 per cada paraula desconeguda que es vol classificar. Excepte que ara els coeficients MFCC obtinguts provenen d'un sol enregistrament (paraula desconeguda).

5. El proper pas és identificar la paraula mitjançant un classificador que en aquest cas consisteix en avaluar el valor de log-likelihood entre la paraula desconeguda i tots els models obtinguts en el punt 3. De tots els valors obtinguts s'escull el valor màxim.
Per obtenir els valors de log-likelihood s'ha utilitzat la funció de Matlab:

```
[P,NLogl(j)]= posterior(Models{j},coeficientsMEL);
```

Aquesta funció retorna la probabilitat que un conjunt de dades no pertany a un model determinat quantificat en el valor *NLogl* (valors invers al logl-likelihood) per tant en el codi Matlab escollirem el valor Mínim entre tots els *NLogl* en lloc del valor Màxim que es mostra en la *figura 3.1*.

En les següents *figures 3.2 i 3.3* podem veure dos exemples del models obtinguts amb la mescla de Gaussianes, les barres blaves corresponen al histograma dels coeficients MFCC dels enregistraments de paraula "Dreta" i "Amunt". El traç vermell correspon a la suma de mescles de Gaussianes de dimensió 1.

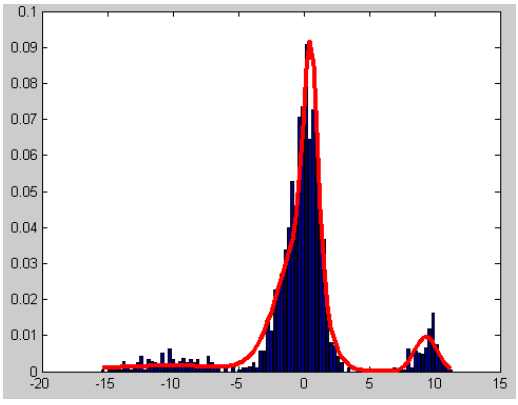


Figura 3.2 Model GMM paraula DRETA

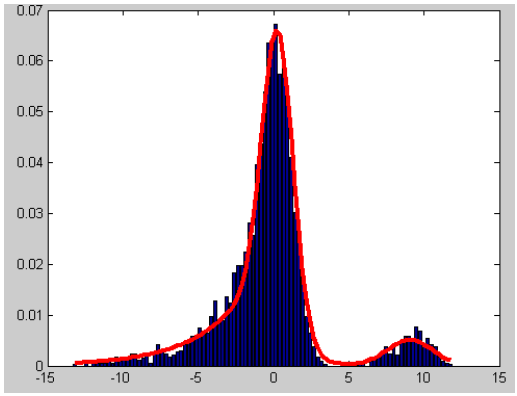


Figura 3.3 Model GMM paraula AMUNT

El classificador utilitzat en aquest reconeixedor està basat en el *likelihood* o probabilitat a posterior de que un conjunt de característiques d'un senyal de veu pertanyin a un model concret. Hi ha altres formes d'implementar el classificador i una d'elles és utilitzar un Xarxa Neuronal, així que també es va procedir a parametritzar una XNA per classificar els paràmetres obtinguts mitjançant GMM.

3.3.2 Reconeixedor basat en Xarxes Neuronals Artificials

Amb aquest reconeixedor s'utilitzen els coeficients MFCC del senyal de veu per entrenar una Xarxa Neuronal que fa la funció de classificador (*veure figura 3.4*).

Per entrenar la xarxa s'ha utilitzat 11 Locutors (5 homes i 6 dones), 5 enregistraments per paraula. En la fase de Test es van utilitzar 10 locutors independents dels locutors utilitzats per entrenar la Xarxa (5 dones i 5 homes).

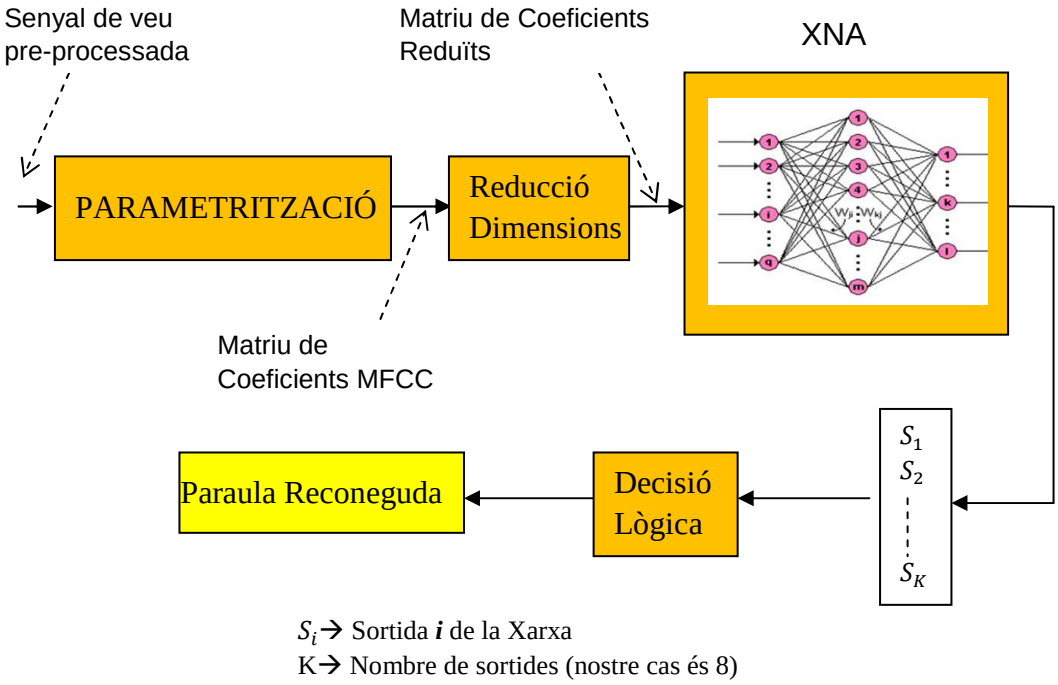


Figura 3.4 Reconeixedor basat en XNA

1 Abans de la parametrització del senyal de veu aquest passa per l'etapa de pre-processament tal com s'ha detallat en l'apartat 2.3

2 En l'etapa de parametrització s'obtenen els coeficients MFCC del senyal, en aquest cas el millor resultat va ser per 20 coeficients MFCC.

3 En l'etapa de reducció de la dimensió de característiques el que es busca es disminuir la càrrega computacional de la xarxa ja que en l'etapa anterior s'han obtingut 20 coeficients per cada trama del senyal de veu de 10 ms. La solució és fer la mitja dels coeficients obtinguts de cada trama, així aconseguim tenir un vector de característiques per cada paraula de dimensió 1x20.

4. La Xarxa Neuronal parametritzada consta dels següents paràmetres:

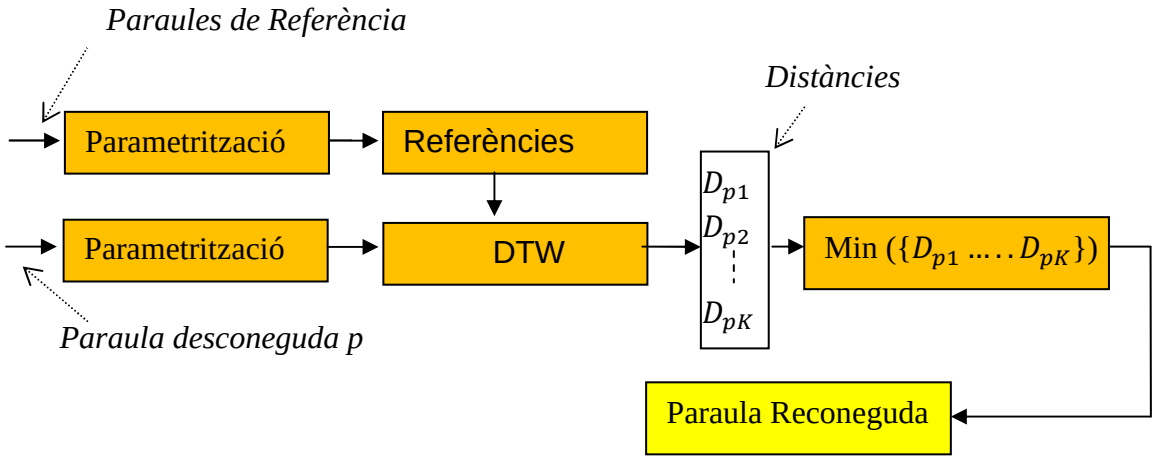
- 20 Entrades (una per cada coeficient MFCC)
- 40 Neurones a la capa oculta
- 8 Sortides (una per cada paraula)
- Màxim de 50000 èpoques
- Funció de capa d'entrada *'tansig'*
- Funció de capa oculta '*tansig*'
- Funció de sortida '*logsig*'
- Entrenament amb funció '*traincgf*' . Algoritme de gradient conjugat en la variació de Fletcher-Reeves.

5. En funció del valor màxim de les sortides es determina quina sortida prendrà el valor lògic 1 i la resta prendran el valor lògic 0. La paraula reconeguda es determina per l'índex que ens dona el valor 1 de les 8 sortides (una sortida per cada paraula).

La desavantatge d'aquest sistema és que la XNA no té en compte el caràcter temporal del senyal de veu i això fa disminuir en gran mesura l'índex d'efectivitat. És a dir, una mateixa paraula pot tenir diverses longituds en funció de la pronunciació ja sigui pel mateix locutor o un altre locutor, aquest factor dificulta molt el reconeixement. Com a avantatge és que un cop entrenada la xarxa és molt ràpid el reconeixement ja que són operacions bàsiques de sumes i multiplicacions.

3.3.3 Reconeixedor basat en l'Alineament Temporal Dinàmic

En aquest tipus de reconeixedor es va utilitzar l'algoritme DTW per obtenir les distancies entre un grup de paraules de Referència i la paraula desconeguda. La distancia mínima determina l'índex de la paraula que és reconeguda.
Es van utilitzar 11 Locutors de referència (5 homes i 6 dones) i 10 Locutors de Test.



D_{pj} → *Distancia entre la Paraula desconeguda p i la Paraula j de Referència*

K → *nombre de referències*

Figura 3.5 Reconeixedor basat en DTW

1. La primera etapa és la de parametrització de les mostres que s'utilitzaran com a referència. Es van utilitzar 12 coeficients MFCC.

2. Un cop obtinguts els paràmetres de les mostres de referència, es calculen els paràmetres MFCC de la paraula desconeguda.

3. Per cada mostra de referència es calcula la distancia euclidiana amb la paraula desconeguda utilitzant l'algoritme DTW, obtenint tantes distàncies com nombre de referències utilitzades.

4. Del vector de distancies s'escull el valor mínim i segons la posició es coneix la paraula que representa.

La principal desavantatge és que hi ha un temps de processament de l'algoritme DTW i per tant com més referències utilitzem el temps total s'incrementa. Per contra com més referències més efectivitat en el reconeixement. En el sistema presentat s'han utilitzat 55 referències per cada paraula i el temps total per reconèixer una paraula va ser de 46 minuts.

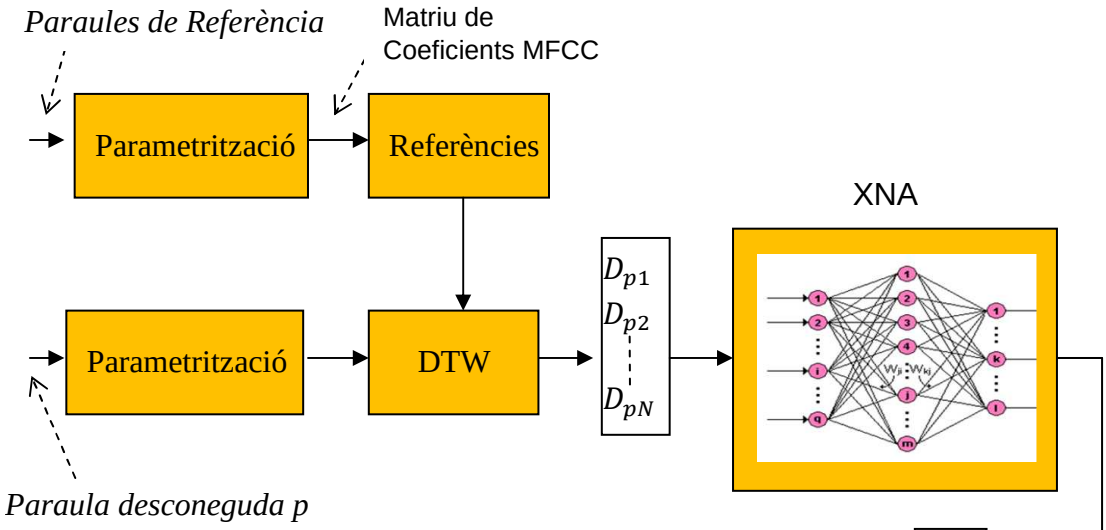
L'avantatge principal com veurem en l'apartat 3.5, és que l'efectivitat és molt millor que en els sistemes presentats en els apartats anteriors, GMM i XNA.

3.3.4 Reconeixedor híbrid DTW i XNA

Els millors resultats dels reconeixedors experimentats van arribar implementant un híbrid mitjançant l'algoritme DTW i una Xarxa Neuronal com a classificador. Aquest reconeixedor aprofita les avantatges dels dos sistemes i com es veurà en l'apartat 3.5 assoleix un índex d'efectivitat força elevat amb locutor independent.

Per aquest reconeixedor s'ha implementat una interfície gràfica que es pot utilitzar per reconèixer paraules pronunciades per locutors independents, és a dir no utilitzats en la fase d'entrenament i tampoc com a registres de referència per l'algoritme DTW.

En definitiva és el reconeixedor utilitzat en la demostració pràctica, amb el qual es poden entrenar diverses xarxes segons les mostres obtingudes, les referències i el nombre de neurones de la capa oculta. Les Xarxes i els seus paràmetres es guarden en fitxers d'extensió *.mat* que posteriorment seran utilitzades per provar el reconeixedor escollint la que es vulgui.



S_i → Sortida ***i*** de la Xarxa
K → Nombre de sortides (nostre cas és 8)
N → Nombre de referències

Figura 3.6 Reconeixedor híbrid DTW-XNA

1. El primer pas després del pre-processament, com tots els reconeixedors és la parametrització o extracció de característiques, al igual que en el reconeixedor basat en DTW s'extreuen els coeficients MFCC per totes les Referències escollides (veure 3.4 Interfície gràfica) i es guarden en una variable del tipus *"Cell Array"*. Els millors resultats es van produir amb 12 coeficients MFCC, sense contar amb l'energia i els coeficients delta i doble delta que també es van incloure (39 coeficients en total)

2. En la fase d'entrenament s'obtenen tots els coeficients MFCC de les mostres escollides per entrenament i també es guarden en una variable *"Cell Array"*

3. Per cada paraula d'entrenament s'obté un vector de distancies amb totes les paraules de Referència, els vectors obtinguts s'acumulen en una matriu de distancies que serà la matriu utilitzada per entrenar la xarxa neuronal.

4. El mateix procediment que l'anterior s'aplica per les mostres de test però aquestes serviran per testejar la xarxa neuronal. Es genera un matriu de test de distancies a partir de les mostres de test gravades o bé a partir d'una mostra gravada a temps real.

5. A la capa de sortida de la xarxa neuronal tenim 8 sortides una per cada paraula del vocabulari. El valor més alt correspondrà a la paraula reconeguda. La Decisió Lògica dependrà d'un valor llindar mínim (paràmetre) per sobre del qual es considera paraula reconeguda i per sota es considera paraula no reconeguda.

La xarxa neuronal i el nombre de mostres de referència es poden parametritzar, de tal forma que els millors resultats obtinguts es van donar utilitzant els valors indicats en la taula. Concretament aquesta configuració va donar un **94,42%** d'efectivitat.

$$Efectivitat = \frac{Nombre\ Total\ Acerts}{Total\ Mostres\ Test} * 100 \qquad \textit{Equació 3.1}$$

Locutors de referència (un home i una dona)	2
Mostres de referència per locutor i paraula	10
Locutors d'entrenament (6 homes i 9 dones)	15
Mostres d'entrenament per locutor i paraula	5
Locutors de test (6 homes i 7 dones)	13
Mostres de test per locutor i paraula	5
Nombre de neurones capa oculta	400
Nombre de neurones d'entrada	160

Taula 3.1 Paràmetres Reconeixedor Híbrid DTW-XNA

3.4 Interfície gràfica

Per gestionar tots els *scripts* i tenir una eina més visual s'ha utilitzat la capacitat de Matlab per crear interfícies d'usuari. El primer és el menú principal (fig.3.7) on es poden cridar la resta de pantalles d'usuari.

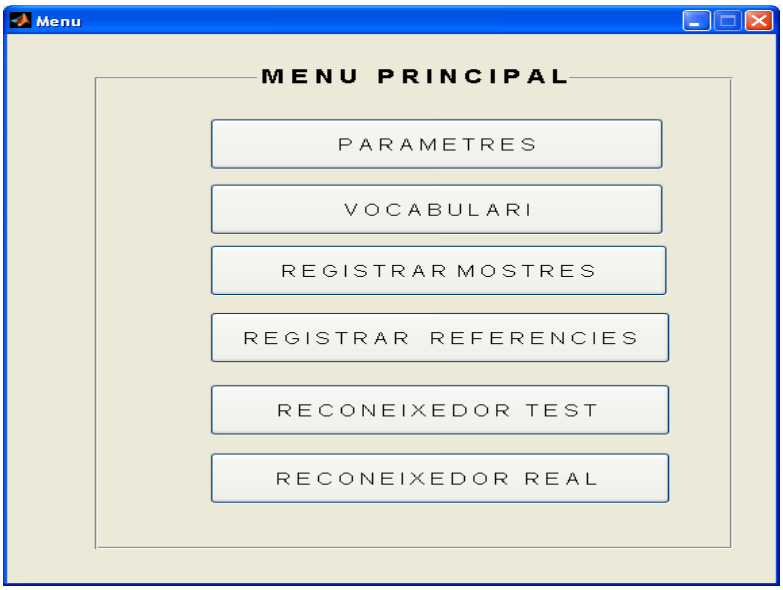


Figura 3.7 Menú Principal (menu.m)

Cal afegir les paraules escollides al vocabulari permès pel reconeixedor amb el botó "VOCABULARI"

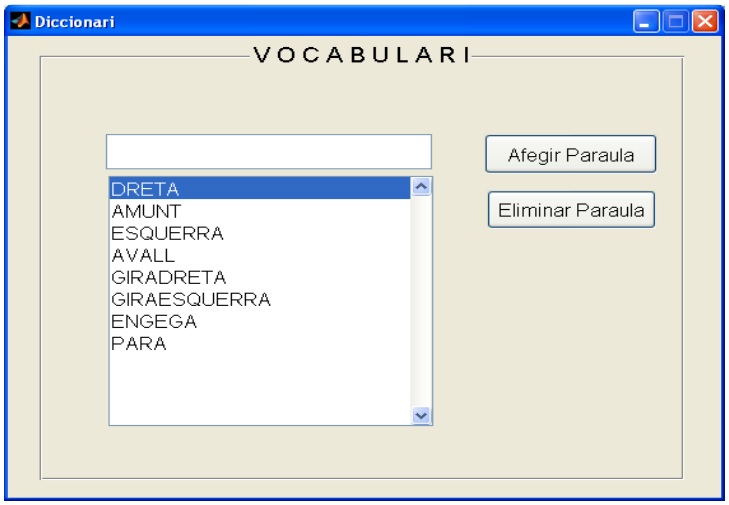


Figura 3.8 Vocabulari (diccionari.m)

El proper pas és cridar la pantalla de paràmetres i omplir els paràmetres que s'utilitzaran per entrenar i testejar el classificador híbrid (DTW-XNA).

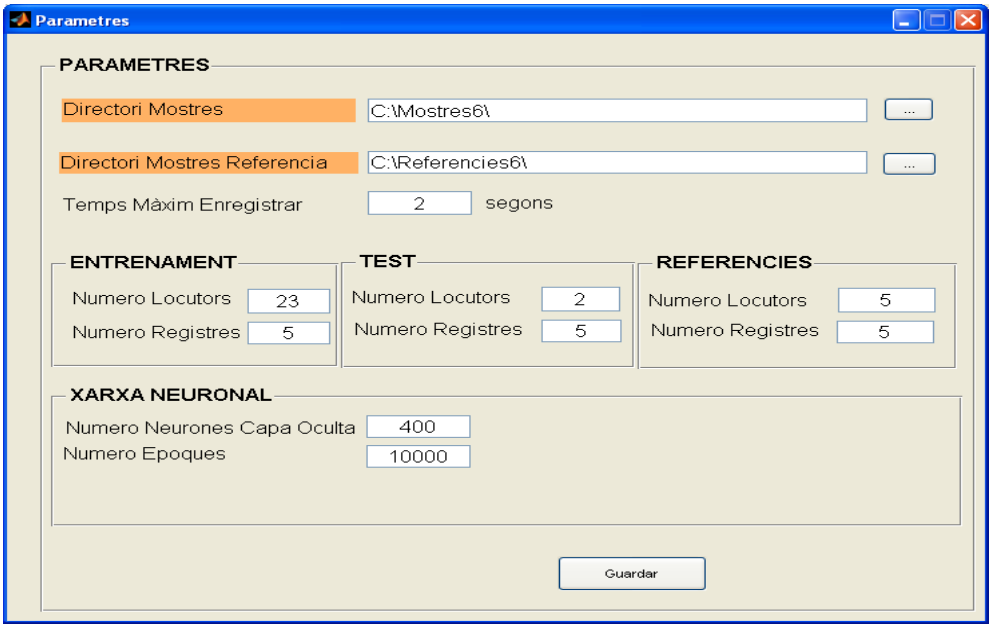


Figura 3.9 Paràmetres (parametres.m)

Per confeccionar el corpus de veus es va utilitzar la interfície de la figura 3.11 amb la qual generem els fitxers d'àudio en format *wav*, amb la nomenclatura següent:

Exemple: C1L3R2H39.wav

C1→ ***'C'*** Indica Classe i el número de classe (en aquest cas primera paraula del vocabulari)

L3→ '***L'*** Indica Locutor i el número de locutor (en aquest cas locutor 3)

R2→ ***'R'*** Indica Registre i el número de registre (en aquest cas el segon registre)

H39→ ***'H'*** Indica que és un Home i l'edat de 39 anys. Per una dona seria amb lletra ***'D'***

Les mostres tant d'entrenament com de test es guarden en un únic directori indicat en la interfície de paràmetres, amb el següent format (*fig. 3.10*) :

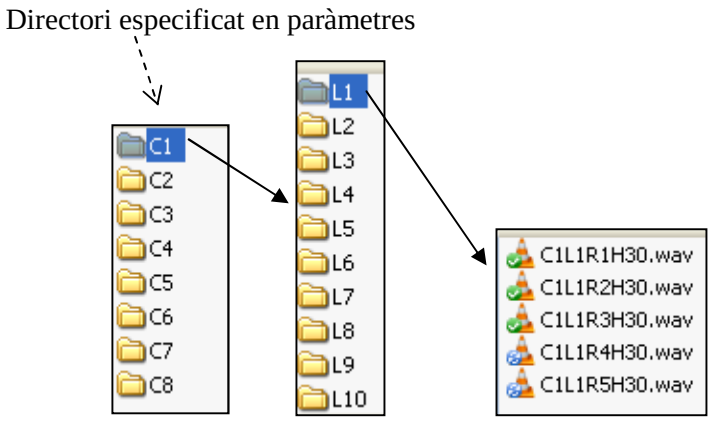


Figura 3.10. Exemple del directori de la base de dades de veus

Els primers locutors corresponen als d'entrenament, per exemple si el número de locutors d'entrenament és 8 el sistema escollirà els 8 primers locutors del directori per entrenament. Els de test seran els següents als d'entrenament, per exemple si el número de locutors de test és 2, els locutors de test seran el L9 i L10 del directori.

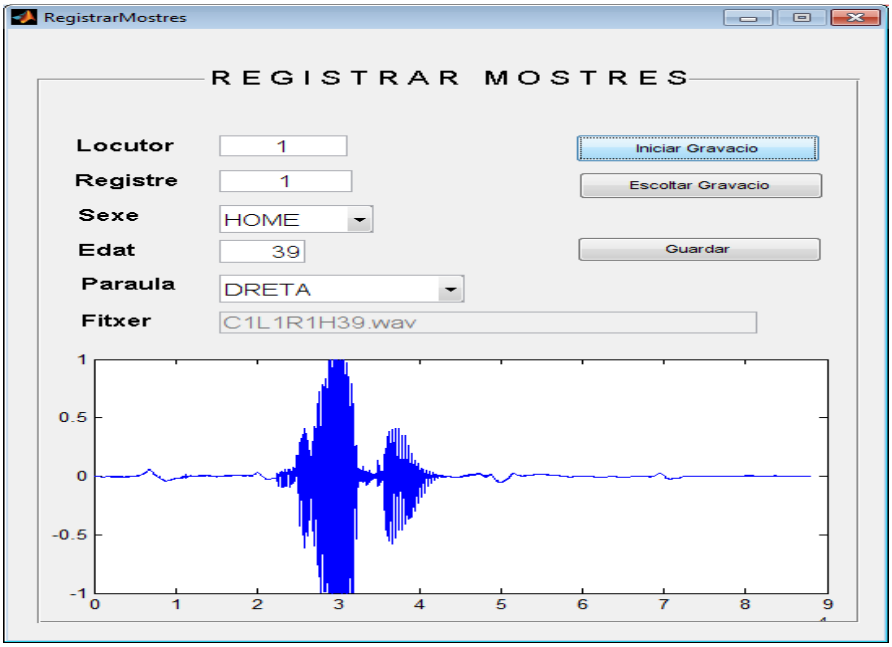


Figura 3.11 Registre de mostres (RegistrarMostres.m)

Per registrar les mostres que s'utilitzaran com a referència s'utilitza la següent interfície (fig. 3.12). La nomenclatura dels fitxers d'àudio guardats és la mateixa que

per les mostres d'entrenament i test, la única diferència és que les referències es guarden en un altre directori segons s'hagi indicat en paràmetres.

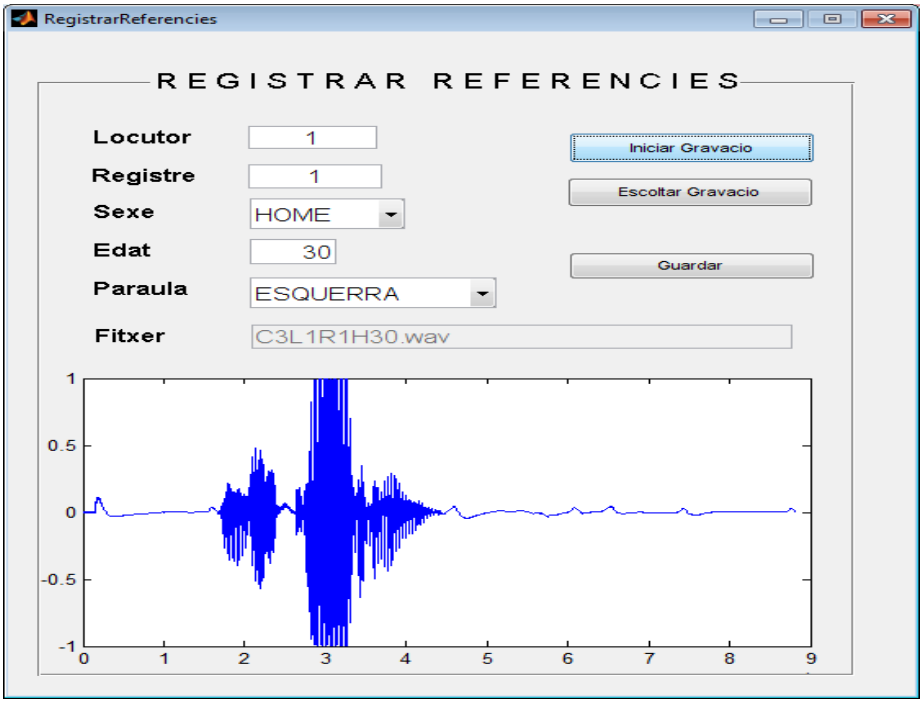


Figura 3.12 Registre de referències (RegistrarReferencies.m)

Per poder entrenar i testejar el reconeixedor segons els paràmetres introduïts en paràmetres, cal utilitzar aquesta interfície (fig. 3.13) , primer cal posar un nom descriptiu de la xarxa en l'apartat d'entrenament, després prémer "Matriu per entrenar" això genera la matriu de distàncies d'entrenament per la xarxa neuronal.

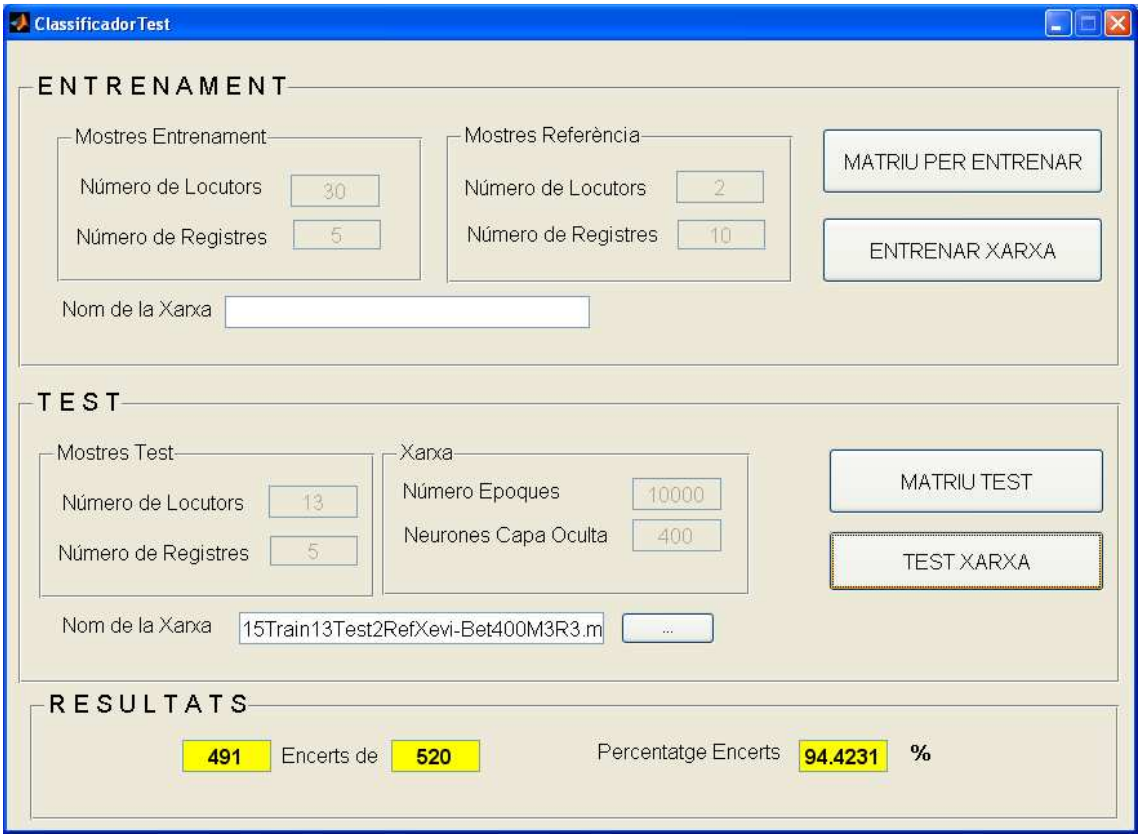


Figura 3.13 Classificador de test (ClassificadorTest.m)

Posteriorment cal prémer "Entrenar Xarxa" per iniciar el procés d'entrenament de la xarxa neuronal (figura 3.13) . Un cop entrenada escollirem el fitxer amb el nom de la xarxa que hem generat a l'entrenament i generem la matriu de distàncies de test amb el botó "Matriu Test", finalment un cop generada la matriu de distàncies de test premem el botó "Test Xarxa" i podrem veure els resultats en la part inferior de "Resultats". El fitxer generat conté tots els paràmetres utilitzats per entrenar la xarxa i ens servirà posteriorment per la interfície del reconeixedor real.

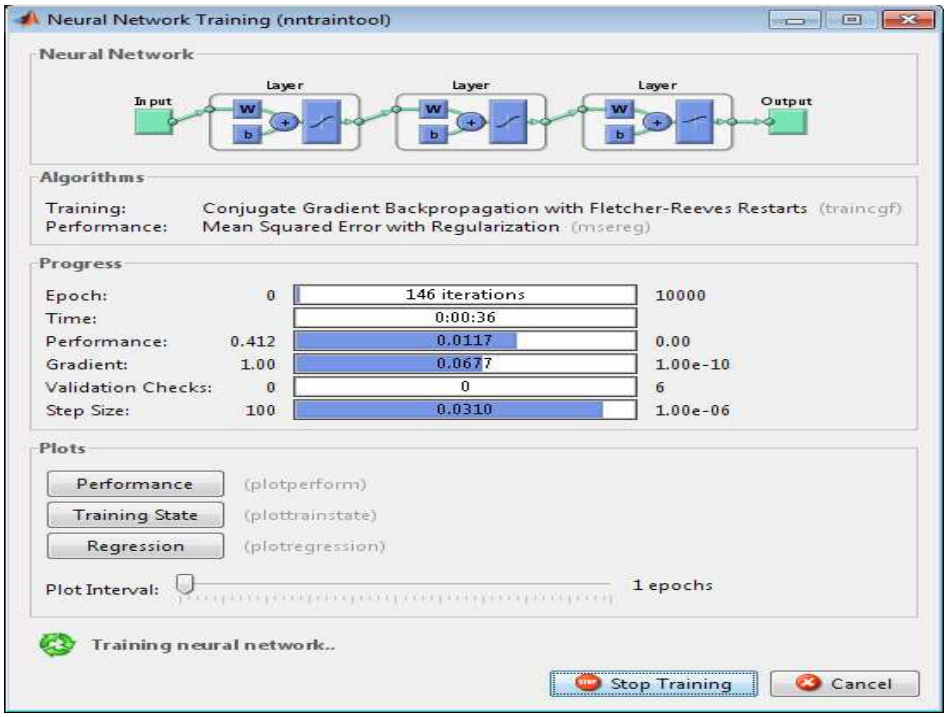


Figura 3.14 Procés d'entrenament de la xarxa neuronal

Per reconèixer paraules una a una en el moment de ser pronunciades utilitzarem la interfície de la figura 3.15. Cal escollir un dels fitxers generats en l'etapa d'entrenament i prémer el botó de "Carregar Paràmetres". En el moment de prémer el botó "Iniciar" el sistema espera a que el locutor pronunciï una paraula indefinidament. En el moment en que es pronuncia una paraula el locutor disposa de 2 segons de duració per pronunciar la paraula.

Per poder ajustar la distància entre el locutor i el micròfon el sistema disposa d'uns marges de seguretat de l'energia del senyal en els quals la captura es considera vàlida, per sota d'un valor d'energia de 300 es considera que el volum és massa dèbil i per tant cal apropar-se més al micròfon o bé pronunciar la paraula més fort. Per contra si el valor de l'energia del senyal és superior a 2000 es considera un excés de volum (saturació del senyal) en aquest cas caldrà allunyar-se del micròfon o bé pronunciar la paraula més fluix.

El camp "Nivell de Decisió" indica el límit pel qual es considera vàlida una paraula o no, és a dir si no hi ha cap sortida de la xarxa neuronal que superi aquest llindar es considera que la paraula no pertany al vocabulari i que no és reconeguda.

Aquest paràmetre es pot modificar en tot moment però per defecte s'inicialitza amb el valor 0.90.

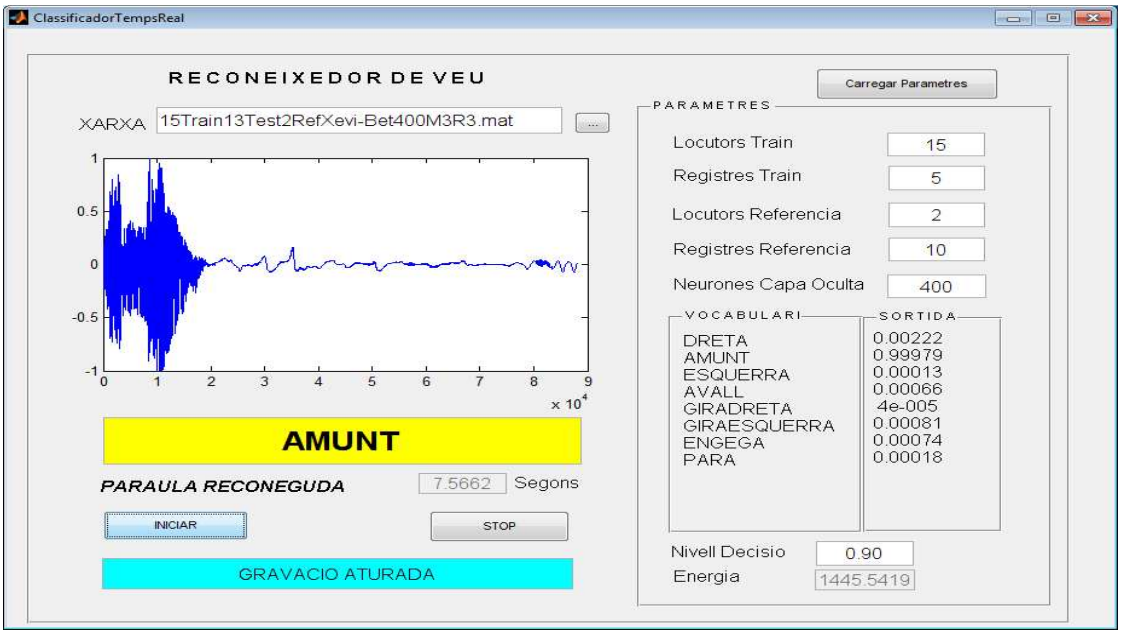


Figura 3.15 Classificador Real (ClassificadorTempsReal.m)

El següent codi inicialitza la tarja de so i es manté a l'espera d'un senyal de veu fins que una tensió d'entrada provinent del micròfon superi el llindar de 0.13V, en aquest moment comença la captura de dades a temps real de la tarja de so. La gravació es mantindrà durant 2 segons des del moment en que comença a capturar dades.

```
%********** Paràmetres Tarja de so ************

AI = analoginput('winsound');
chan = addchannel(AI,1);
duracio = 2;
set(AI,'SampleRate',fs);
ActualRate = get(AI,'SampleRate');
set(AI,'SamplesPerTrigger',duracio*ActualRate);
set(AI,'TriggerType','software');
set(AI,'TriggerCondition','rising');
set(AI,'TriggerConditionValue',0.13);   % Activació del micròfon
set(AI,'TriggerChannel',AI.Channel(1));
set(AI,'TriggerDelay',-0.25);
set(AI,'TimeOut',600);
preview = duracio*ActualRate/100;
set(gcf,'doublebuffer','on');
SenyalVeu =[];
start(AI)
```

En ocasions en donar un petit cop al micròfon és produeix un senyal de soroll de curta durada, que és detectat pel sistema com una paraula, normalment amb la paraula "dreta". Per evitar que el reconeixedor analitzi senyals de soroll de curta durada es calcula la taxa de creuaments per zero (funció ***zcr***) i si el valor obtingut és inferior a un llindar (0.04) es considera soroll i per tant no és analitzat donant un missatge en pantalla com a paraula "No Reconeguda".

```
function zcr=zcr(x,dur)
% funció zcr=zcr(x,dur) : computa la taxa de creuament per zero
% x: dades d'entrada
% dur: duració del senyal d'entrada

[nf,len]=size(x);
zcr=sum(0.5*abs(sign(x(:,2:len))-sign(x(:,1:len-1))))/dur;
```

3.5 Resultats amb els diferents classificadors

- Els millors resultats obtinguts pel reconeixedor GMM de l'apartat 3.3.1 son els següents:
  - 440 mostres d'entrenament (11 locutors)
  - 400 mostres de test (10 locutors)
  - 308 Encerts de 400 enregistraments
  - 39 coeficients MFCC reduïts a 12 per WMFCC
  - Efectivitat del 77%

Com es pot veure en la taula 3.2 de confusió que indica el nombre d'encerts per paraula i els errors (paraules reconegudes equivocadament), la paraula "Amunt" té un 100% d'encerts mentre que la paraula "Para" té només un 48% d'encerts i la majoria d'errades es produeixen al identificar erròniament la paraula "Avall" 19 vegades. També hi ha errors entre les paraules "Esquerra" i "Giraesquerra" o bé "Dreta" i "Giradreta".

77%	DRETA	AMUNT	ESQUERRA	AVALL	GIRADRETA	GIRAESQUERRA	ENGEGA	PARA
DRETA	38	0	1	0	10	0	1	0
AMUNT	0	50	0	0	0	0	0	0
ESQUERRA	1	0	36	0	0	13	0	0
AVALL	0	5	0	39	0	4	1	1
GIRADRETA	3	0	0	0	39	3	5	0
GIRAESQUERRA	0	0	0	0	3	45	2	0
ENGEGA	0	0	0	0	8	5	37	0
PARA	1	4	0	19	0	2	0	24

Taula 3.2 Matriu confusió reconeixedor GMM

- Pel que fa al reconeixedor de l'apartat 3.3.2 basat en una Xarxa Neuronal son els següents:
  - 440 mostres d'entrenament (11 locutors)
  - 400 mostres de test (10 locutors)
  - 20 coeficients MFCC
  - 8 neurones d'entrada
  - 30 neurones capa oculta
  - 266 Encerts de 400 enregistraments
  - Efectivitat del 66,5%

66,50%	DRETA	AMUNT	ESQUERRA	AVALL	GIRADRETA	GIRAESQUERRA	ENGEGA	PARA
DRETA	34	0	0	0	15	0	1	0
AMUNT	0	48	0	1	0	1	0	0
ESQUERRA	1	0	39	0	0	10	0	0
AVALL	2	3	0	32	0	0	0	13
GIRADRETA	5	0	0	0	31	4	10	0
GIRAESQUERRA	0	0	5	0	10	33	2	0
ENGEGA	1	0	3	0	13	7	26	0
PARA	5	0	2	20	0	0	0	23

Taula 3.3 Matriu confusió reconeixedor XNA

A la *taula 3.3* es pot apreciar que el pitjor resultat es dona per la paraula "Para" que es reconeguda com la paraula "Avall", probablement per tenir la mateixa longitud de fonemes i les mateixes vocals. El resultat de 66,5% és el pitjor de tots, es van provar diferents paràmetres per la Xarxa, però en cap cas es va superar la cota de 66,5%. Una

possible explicació podria ser deguda a la reducció de dimensions en les dades d'entrada de la xarxa utilitzant la mitjana de coeficients. Existeix altres mètodes com Anàlisis de Components Principals (PCA) que potser donarien millors resultats, però de totes maneres, el principal problema és que cada enregistrament d'una mateixa paraula té diferent longitud degut a varis factors i que aquest problema queda resolt utilitzant l'algoritme DTW *(veure apartat 2.5.2)*

- En el cas del reconeixedor DTW *(3.3.3)* l'índex d'efectivitat augmenta considerablement però repercuteix en el rendiment ja que l'algoritme DTW és costós computacionalment.

  - 440 mostres d'entrenament (11 locutors)
  - 400 mostres de test (10 locutors)
  - 373 Encerts de 400 enregistraments
  - 39 coeficients MFCC reduïts a 12 per WMFCC
  - Efectivitat del 93.25%

Cal esmentar que s'han utilitzat totes les mostres d'entrenament com a mostres de referència per tant s'han calculat 440*400 = 176.000 distàncies. A més a més per assolir el 93,25% d'efectivitat es van utilitzar trames de veu de 512 mostres i 256 de solapament, la qual cosa significa un nombre elevat de coeficients per paraula. Així que per obtenir els resultats de la *taula 3.4* ha calgut 16,46 hores de processament, és a dir amb aquests paràmetres l'algoritme DTW necessita 3 segons per calcular la distància entre dues paraules (també dependrà de la velocitat del processador i altres factors de l'ordenador on es fan els càlculs)

93,25%	DRETA	AMUNT	ESQUERRA	AVALL	GIRADRETA	GIRAESQUERRA	ENGEGA	PARA
DRETA	46	0	1	0	0	0	1	2
AMUNT	0	39	0	9	0	0	0	2
ESQUERRA	0	0	48	0	0	0	2	0
AVALL	0	0	0	50	0	0	0	0
GIRADRETA	0	0	0	0	46	0	4	0
GIRAESQUERRA	0	0	0	0	0	50	0	0
ENGEGA	1	0	0	0	3	0	46	0
PARA	2	0	0	0	0	0	0	48

Taula 3.4 Matriu confusió reconeixedor DTW

- Per últim el millor resultat *(Taula 3.5)* es va ser obtingut amb el reconeixedor DTW-XNA de l'apartat 3.3.4, aconseguint dos objectius que sigui suficientment ràpid per fer una demostració de reconeixement real i que sigui molt efectiu.

  - 2 Locutors de referència (un home i una dona)
  - 10 Mostres de referència per locutor i paraula
  - 15 Locutors d'entrenament (6 homes i 9 dones)
  - 5 Mostres d'entrenament per locutor i paraula
  - 13 Locutors de test (6 homes i 7 dones)
  - 5 Mostres de test per locutor i paraula
  - 400 neurones capa oculta
  - 160 neurones d'entrada (una per cada enregistrament de referència)
  - 491 Encerts de 520 enregistraments
  - **94,42 %** d'efectivitat

94,42%	DRETA	AMUNT	ESQUERRA	AVALL	GIRADRETA	GIRAESQUERRA	ENGEGA	PARA
DRETA	60	2	2	0	0	0	0	1
AMUNT	0	56	3	3	0	1	0	2
ESQUERRA	3	0	57	0	0	0	5	0
AVALL	0	0	2	63	0	0	0	0
GIRADRETA	0	0	0	0	64	1	0	0
GIRAESQUERRA	0	0	0	0	0	65	0	0
ENGEGA	0	1	0	0	1	0	63	0
PARA	1	0	0	1	0	0	0	63

Taula 3.5 Matriu confusió reconeixedor híbrid DTW-XNA

S'han fet proves amb tres locutors de referència (1 home i dues dones), també amb només un locutor de referència (1 home i després 1 dona), però cap de les combinacions provades supera la configuració d'utilitzar com a referències dos locutors, un home i una dona. Cal destacar que el nombre òptim de neurones de la capa oculta es donava amb 400 neurones, això s'ha trobat de forma empírica provant diferents valors.

CONCLUSIONS

4. CONCLUSIONS

4.1 Resultats i conclusions

De forma clara es pot afirmar que s'han assolit els objectius plantejats en la introducció del treball i de forma satisfactòria. Després de varis prototips i utilitzant diverses tècniques de reconeixement de veu s'ha trobat la metodologia adient per obtenir un sistema robust amb un índex d'efectivitat del 94% i suficientment ràpid per fer una demostració real del reconeixement de paraules aïllades. El més important és que s'ha aconseguit que el sistema sigui estrictament independent de locutor, cosa que era la part més difícil d'obtenir ja que la majoria de desenvolupaments trobats a Internet eren dependents de locutor. També cal destacar la integració de les Xarxes Neuronals com a part fonamental del classificador en el model híbrid DTW-XNA.

Les dificultats en assolir un 94% d'efectivitat han estat molt grans i per la meva part no estaven previstes, la gran quantitat de literatura associada al tema de reconeixement de veu, la multitud de mètodes estudiats i sobretot la gran quantitat de paràmetres que intervenen al llarg de totes les etapes del reconeixedor, van repercutir de forma negativa en el temps de desenvolupament.

Un altre dificultat afegida al desenvolupament, va ser l'elaboració del corpus de veus en condicions poc favorables pel reconeixement, ja que no es disposava dels mitjans adequats com per exemple una càmara anecoica, un micròfon d'alta gama professional o un equip de so professional. Per tant les mostres obtingudes incorporen elements de soroll, saturacions de micròfon i altres inconvenients que dificulten molt el reconeixement, malgrat això aquests elements estan inclosos en el 94% d'efectivitat. Així doncs tenint en compte que a dia d'avui no existeix cap reconeixedor amb una efectivitat del 100% podem considerar un gran èxit el reconeixedor plantejat.

4.2 Possibles millores

Un dels aspectes a millorar és l'adquisició del senyal d'àudio, tot i que el reconeixedor ha mostrat tenir una bona efectivitat i robustesa, val a dir que utilitzar l'entrada "mic" d'una tarja de so d'un ordenador portàtil no ofereix garanties de qualitat, encara que tinguem el millor micròfon del mercat, aquests no estan pensats per ser utilitzats amb ordenadors. A banda l'entrada "*mic"* pròpia dels micròfons es un senyal dèbil que cal amplificar (sobre uns 26db) per adequar-la a la resta de senyals i això fa que si hi ha soroll a l'entrada aquest també s'amplifiqui. Així que la millor solució seria utilitzar un micròfon professional amb un amplificador (previ) o consola de mescles a l'entrada de la tarja de so "*Line-in"* que és un senyal més potent i no necessita amplificació interna.

El prototip presentat en la demostració pràctica va ser dissenyat per comprovar els resultats del reconeixement en real però no per fer un reconeixement immediat de cada paraula ja que el Matlab no ofereix una velocitat de processament prou elevada al tractar-se

d'un llenguatge interpretat, per tant un següent pas al prototipus presentat seria escriure el codi en llenguatge C que és un llenguatge compilat i per tant molt més ràpid. Almenys com a mínim s'hauria d'escriure en C l'algoritme DTW que és on es perd més temps de processament.

Un altre aspecte a millorar podria ser la robustesa en presència de soroll extern ocasionat per dispositius elèctrics, com ara ventiladors, l'aire condicionat, llums fluorescents, etc. També s'ha detectat que és sensible a les ressonàncies del lloc on està situat degut a reflexions i retards en el temps de retorn del senyal original d'una superfície reflectant com una paret o una finestra. Per això una solució encara que no el 100% efectiva seria dissenyar filtres específics per reduir el soroll o bé utilitzar un micròfon molt més direccional a la font d'àudio.

5. BIBLIOGRAFIA

5. BIBLIOGRAFIA

Llibres

- ISASI VIÑUELA, Isasi Pedro - M.GALVAN LEON, Inés
  Redes de Neuronas Artificiales Un Enfoque Práctico
  Pearson – Prentice Hall
- FAUNDEZ ZANUY, Marcos
  Tratamiento digital de voz e imagen
  Marcombo, S.A. (2000)
- R.DELLER, Jr. John - H.L.HANSEN, John - G.PROAKIS, John
  Discrete-Time Processing of Speech Signals
  IEEE Press Editorial Board (2003)

Adreces d'Internet

- VIQUIPÈDIA l'enciclopèdia lliure. *Decodificador acústic fonètic*

http://ca.wikipedia.org/wiki/Decodificador_ac%C3%BAstic_fon%C3%A8tic
(consulta: 10-07-2012)

- VIQUIPÈDIA l'enciclopèdia lliure. *Aparell fonador*
  http://ca.wikipedia.org/wiki/Aparell_fonador
  (Consulta: 10-07-2012)
- VIQUIPÈDIA l'enciclopèdia lliure. *Aparell auditiu*
  http://ca.wikipedia.org/wiki/Aparell_auditiu
  (Consulta: 10-07-2012)
- VIQUILLIBRES Llibres lliures per un món lliure. *So/Audició i sistema auditiu*
  http://ca.wikibooks.org/wiki/So/Audici%C3%B3_i_sistema_auditiu
  (Consulta: 10-07-2012)
- BROOKES, Mike. Department of Electrical & Electronic Engineering, Imperial College London. *VOICEBOX: Speech Processing Toolbox for MATLAB*
  http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (Consulta: 09-05-2012)
- MARTINEZ OLALLA, Rafael. Universidad Politécnica de Madrid.
  El Sistema de Producción de Voz
  http://tamarisco.datsi.fi.upm.es/ASIGNATURAS/TDSV/Modelo_Produccion.pdf
  (Consulta: 10-05-2012)
- ELAMIN ELNIMA ELGASIM. King Saud University. College of Computer and Information Sciences. *Relative distance vector neural network (RDVNN) model: a hybrid approach to speech recognition.*
  http://repository.ksu.edu.sa/jspui/handle/123456789/3293
  (Consulta: 15-07-2012)
- V.CHAPANERI, Santosh. Department of Electronics and Telecommunication Engineering, Universtiy of Mumbai. *Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping*.
  http://research.ijcaonline.org/volume40/number3/pxc3877167.pdf
  (Consulta:15-07-20212)
- LUH TAN, Chin and JANTAN, Adznan. Faculty of Engineering University Putra Malaysia. *Digit Recognition using neural networks*.
  http://mjcs.fsktm.um.edu.my/document.aspx?FileName=308.pdf
  (Consulta: 10-05-2012)
- HUMBERTO PECH CARMONA, Jaime. Instituto politécnico nacional centro de investigación en computación, México D.F. Junio del 2006. *Desarrollo de un sistema de reconocimiento de voz para el control de dispositivos utilizando mixturas gaussianas.*

http://itzamna.bnct.ipn.mx:8080/dspace/handle/123456789/1427?mode=full
(Consulta:10-06-2012)

- NING, Daryl. MATLAB Digest – January 2010. *Developing an Isolated Word Recognition System in Matlab.*
  http://www.mathworks.es/company/newsletters/digest/2010/jan/word-recognition-system-matlab.html (Consulta:20-06-2012)
- BULBULLER, Gokhan. Naval Postgraduate School Monterey, California.
  Recognition of in-ear microphone speech data using multi-layer neural networks.
  http://www.dtic.mil/dtic/tr/fulltext/u2/a445459.pdf
  (Consulta: 20-06-2012)
- PRIETO LABRADOR, Enrique. Universidad Carlos III de Madrid. *Estudio comparativo de parámetros espectrales para clasificación de audio.*
  http://e-archivo.uc3m.es/bitstream/10016/5384/1/PFC_Enrique_Prieto_Labrador.pdf
  (Consulta: 25-06-2012)
- The MathWorks, Inc. *Data Acquisition Toolbox. User's Guide*
  http://www.mathworks.com/help/pdf_doc/daq/daqug.pdf
  (Consulta: 10-07-2012)