REPUBLIKA E SHQIPËRISË
UNIVERSITETI POLITEKNIK I TIRANËS
FAKULTETI I INXHINIERISË MATEMATIKE DHE INXHINIERISË FIZIKE
DEPARTAMENTI INXHINIERISË MATEMATIKE

Për marrjen e gradës

“ Doktor”

M.Sc. ERVENILA MUSTA (Xhaferraj)

DISERTACION

METODA PËR VLERËSIMIN E ZHVILLIMIT SPEKTRAL TË NJOHJES SË FJALËS NËN EFEKTIN E ZHURMAVE

(Studimi i një modeli për njohjen e fjalës për gjuhën shqipe)

Udhëheqës Shkencor : Prof As LIGOR NIKOLLA

Tiranë 2016

REPUBLIKA E SHQIPËRISË
UNIVERSITETI POLITEKNIK I TIRANËS
FAKULTETI I INXHINIERISË MATEMATIKE DHE INXHINIERISË FIZIKE
DEPARTAMENTI INXHINIERISË MATEMATIKE

DISERTACION

i paraqitur nga

M.Sc. Ervenila Musta(Xhaferraj)

Për marrjen e gradës shkencore

"DOKTOR"

në Inxhinieri Matematike

Tema : Metoda për vlerësimin e zhvillimit spektral të njohjes së fjalës nën efektin e zhurmave

" **Studimi i një modeli për njohjen e fjalës për gjuhën shqipe** "

Udhëheqës Shkencor : Prof As LIGOR NIKOLLA

Mbrohet më datë ----------para jurisë

1.----------Kryetar

2.----------Anëtar

3.----------Anëtar

4.----------Anëtar

Familjes time

Prindërve të mi

“ Ne të gjithë besojmë se një shkencë e fjalës është e mundur ,pavarësisht mungesës në këtë fushë të njerëzve që sillen si shkencëtarë dhe rezultateve që duken si shkencë .Shumica njohëse sillen jo si shkenctarë por si shpikës të çmëndur ose inxhinierë të pasigurt….

Përmbajtja

Figura

Abstrakt

Pas komunikimit përmes gjesteve dhe mimikës, të folurit është forma më e vjetër e komunikimit ndërnjerëzor. Për këtë, dhe arsye të tjera, edhe idetë për gjenerimin artificial të të folurit, nga format e tjera komunikuese janë shumë shumë të hershme.

Përpjekjet e para janë të formave të ndryshme, duke filluar nga ato mekanike të shumë shekujve më parë, e deri te zgjidhjet softuerike të dekadave të fundit.

Në kuadër të sintezës së të folurit, pjesë shumë me interes është edhe shndërrimi i të folurit në tekst, apo si do të mund të thuhej leximi automatik i të folurit nga ana e paisjeve elektronike. Tani, është e vetkuptueshme që paisja e tillë elektronike është kompjuteri, por duke përfshirë gjithnjë e më tepër edhe paisjet tjera komunikuese elektronike.

Aplikimet e teknologjisë modern të fjalës janë duke u përdorur gjithnjë e më shumë në mjedise të zhurmshme si në makinë ,në rrugë, në restorante etj.Zhurma e ambjentit e cila në mënyre tipike supozohet që të jetë shtesë ,i corordit mostrat e zërit .Rrjedhimisht nxjerrja e informacionit spectral të fjalës bëhet më e vështirë, për shëmbull në kodimin e të folurit dhe në njohjen e të folurit.Ka shumë modele të studimeve të mëparshme që tregojnë se modeli AR ,sic është ai konvencional ,është i ndjeshëm ndaj zhurmës.

Ky punim fillimisht synon studimin e metodave të avancuara të parashikimit linear për vlerësimin e mbështjellëses spektrale në analizimin e zërit ,sintetizimin e fjalës ,në njohjen e fjalës dhe folësit .Këto janë modele autoregresive të cilat përcaktohen duke përdorur peshën e përkohëshme të katrorit të gabimit parashikues, ose të mbetjes ,në llogaritjen e koeficientëve optimal të filtrit.

Si rast praktik i metodave të avancuara të parashikimit linear, për të treguar përmirësimin dhe rolin e tyre në procesin e nxjerrjes së tipareve apo karakteristikave,në këtë punim , është përdorimi i sistemit të njohjes së fjalës, ASR .U pa e nevojshme pasi sistemi i njohjes së fjalës ASR bëhet problematik nëkushtet e mospërputhjes së njohjes së sinjalit nga korruptimi i zhurmave të ambjentit apo nëkanal që nuk kanë qënë të pranishme gjatë trajnimit .

Punimi gjithashtu bën përpjekje që duke patur parasysh rregullat themelore të gjuhës shqipe si të : alfabetit ,gramatikës ,fonetikës dhe rregullat bazë të shkrim –leximit të hedhi hapat për krijimin e një modelitë një sistemi të tillë të shndërrimit të të folurit në tekst për gjuhën shqipe.

Në punim përfshihet pjesa për përpunimin dixhital të sinjalit që ka të bëjë me përpunimet në sinjalin e dixhitalizuar para se të gjenerohet zëri ,me qëllim të gjenerimit të kuptueshmërisë dhe natyrshmërisë së të folurit.

Gjithashtu përfshihet ndërtimi i një database në gjuhën shqipe që do të përdoret për aplikim .

Së fundmi në punim përfshihet një metodë optimizimi për permiresimin e sistemit të njohjes së fjlalës në fazën e nxjerrjes së tipareve apo karakteristikave . Tiparet më të njohura sot dhe të përdorura për njohjen e të folurit janë koeficientët cepstral të frekuencave Mel (MFCC) të cilat janë marrë fillimisht nga transformimi Furie me kohë të shkurtër (STFT) dhe pastaj analiza e bankave të filtrave në spektrin e frekuencave të shkallës Mel.Për të zgjidhur çështjen e standartizimit të brezit të filtrave dhe numrit të filtrave eshte përdorur teknika e optimizimit të tufës së grimcave ,PSO, për të optimizuar korficientët MFCC.

Duke përdorur këtë teknikë optimizimi ,është realizuar një përmirësim i njohjes së të folurit me anë të rrjetave neural artificial ANN dhe PSO .Domethënë njohja e fjalës është realizuar me anë të kombinimit të nxjerrjes së tipareve apo karakteristikave të PSO dhe rrjetës natyrale artificial ANN.

Abstract

Besides communication through gesture and mimicry, speech is the oldest form of human communication. Because of this and other reasons, the idea of synthesis artificial speech from other forms of communication is very old.

Initial efforts were of different forms, beginning with the mechanical ones from centuries ago, up to software solutions of last decades.

In the context of speech synthesis, a very interesting part is the conversion of written text into speech, or as it may be referred to as automatic reading of texts by electronic devices. It is known today that such electronic device is a computer, but increasingly involving other electronic communication equipment.

Modern applications of speech technology are being used more and more in noisy environments such as in the car, on the street, in restaurants etc.Noise of the environment which are typically assumed to be additional,disorientate the voice samples.Consequently ,extraction of spectral information of speech becomes more difficult, for example in the speech encoding and speech recognition.There are many models of previous studies showing that AR model, as is the conventional, it is sensitive to noise.

This paper aims to study initially advanced forecasting methods to assess the envelopment linear spectral analyzing sound, speech synthesis, speech recognition and speaker. These are autoregression models which are determined using the weight of the interim square forecast error, or residue, in calculating the optimal filter coefficients.

As a practical case of advanced forecasting methods linear to show improvement and their role in the process of extraction of features or characteristics, in this paper, is the use of speech recognition system, ASR. It was deemed necessary after the speech recognition system ASR become problematic in terms of mismatch recognition signal from noise corruption of the environment or the channel that were not present during the training.

The paper also makes effort considering the basic rules of the Albanian language as: alphabet, grammar, phonetics and basic rules of writing -reading to take steps to create a model of such a system of converting speech into text language Albanian.

Included in the paper for digital signal processing that has to do with the signal processing of digitized sound generated before, in order to generate understanding and simplicity of speech.

Also includes the construction of a database on Albanian language to be used for application.

Finally the paper included an optimization method for improving the system of recognition fjlales phase extraction features or characteristics. Features the most popular today and used for speech recognition are coefficients cepstral frequency Mel (MFCC) which are initially received by the transformation Furie short-time (STFT) and then analyzes the banks of filters in the frequency spectrum of scale Mel.For resolve the issue of standardization of band filters and filters the number of optimization techniques is used to herd particles, PSO, to optimize korficientet MFCC.

Using this technique optimization, it realized an improvement of speech recognition through artificial neural networks and PSO.Speech recognition is performed by the combination of issuing PSO attributes or characteristics of natural and artificial ANN network.

Hyrje

Fjala është një valë komplese e informacionit të dendur akustik.Një sinjal i fjalës është një tingull i formuar nga njerëzit për të komunikuar .Si të gjithë zërat apo tingujt e tjerë ai është një turbullim i presionit atmosferik që udhëton në formën e një vale zanore.Për të transportuar fjalën mbi rrjetet e komunikimit ,sinjali fjalës duhet të transformohet nga një fjalë zanore në njësinjal elektrik .Kjo zakonisht është bërë me një mikrofon dhe rezultati është një sinjal elektrik ku tensioni ndryshon me presionin e sinjalit të fjalës.

Idetë për gjenerimin artificial të të folurit, nga format e tjera komunikuese janë shumë shumë të hershme.

Përpjekjet e para janë të formave të ndryshme, duke filluar nga ato mekanike të shumë shekujve më parë, e deri te zgjidhjet softuerike të dekadave të fundit.

Që nga parahistoria e njeriut e deri tek mediat e reja të së ardhmes ,komunikimi i të folurit ka qënë dhe do të jetë mënyra dominuese e lidhjes sociale dhe shkëmbimit të informacionit të njeriut .Përvec ndërveprimit njeri-njeri ,kjo preferencë e njeriut për komunikim në gjuhën e folur gjen një pasqyrim në lidhjen apo ndërveprimin njeri-makinë po aq mirë.Projektimi i një makine që imiton sjelljen e njeriut ,në vecanti aftësinë e të folurit të natyrshëm ,dhe duke iu përgjigjur sic duhet gjuhës së folur ,ka intrigruar inxhinierët dhe shkencëtarët për shekuj me radhë. Homer W.Dudley ishte inxhinier elektronik dhe akustik pionier në këtë fushë duke krijuar zërin e parë elektronik të sintetizuar për Bell Labs në vitin 1930 duke cuar në zhvillimin e një metode të dërgimit të sigurt të trasmetimeve zanore gjatë luftës së dytë botërore .Përparimi me imadh ka ndodhur gati 4 dekada më para me futjen algoritmit të maksimizimit të pritje (Expectation Maximization (EM)) për trajnimin e modeleve të fshehura të Markovit (HMMs).Nëpërmjet algoritmit EM u bë i mundur zhvillimi i sistemit të njohjes së fjalës për detyra të botës reale duke përdorur pasurinë e modeleve të përziera Gausiane (GMM),për të paraqitur lidhjen midis inputit akustik dhe gjëndjeve HMM. Në këtë sistem inputi akustik është krijuar duke lidhur koeficientët cepstral të frekuencave Mel (Mel frequency cepstral coefficients),llogaritjet nga formavalore te papërpunuara dhe diferencat kohore të rendit të parë dhe të dytë .Ky parapërpunim i sinjalit hyrës apo input është projektuar që të hiqet dorë nga sasi të mëdha të informacionit që konsiderohet i parëndësishëm.

Fusha automatike e njohjes së fjalës ASR ,shpërtheu në dekadat e fundit ,pasi njerëzit priren të jenë gjithnjë e më shumë të zënë dhe të kujdesen për pajisjet hands-free dhe eyes-free.Objekti i ASR është kapja e një sinjali akustik të paraqitjes së fjalës dhe përcaktimi i fjalëve që fliten apo që janë parathënë duke përputhur modelin .Për ta bërë këtë, një grup i modeleve akustike dhe gjuhësore ruhen në një bazë të dhënash kompjuterike ,që përfaqësojnë modelet aktuale. Këto modele që rezultojnë pas trajnimit, krahasohen me sinjalet e kapura.

Ky punim përshkruan teknika të përpunimit dixhital të sinjalit për vlerësimin dhe analizimin e sinjalit të burimit të zërit.Sipas teorisë së filtrit- burim të prodhimit të fjalës (Fant 1970)fjala e shprehur mund të ndahet në një sinjal burim që krijohet si rezultat i vibracionit të një palosjeje vokale dhe përgjigjes së filtrit të traktit vokal.Vlerësimi i densitetit spektral të fuqisë së sinjaleve me kohë diskrete është një hulumtim që është studiuar në disa disiplina si dhe në atë të përpunimit të fjalës.Përvec teknikave konvecionale që përdorin transformimin Furie ,një përdorim të gjërë për vlerësimin e spektrit, zënë metodat qebazohen ne modele te parashikimit linear te sinjalit ,te njohura si modele autoregresive .Këto teknika përdoren për të paraqitur informacionin spektral të sinjaleve me kohë diskrete në formë parametrike duke përdorur filtrat dixhital që kanë një strukturë all-pole,në të cilin funksioni i transferimit të Z-domenit përmban vetëm pole.Termi parashikim linear i referohet formulimit të domenit kohë i cili shërben si bazë në optimizimin e koefiçientëve të filtrit all-pole.Cdo mostër e sinjalit në domenin kohë supozohet të shprehet si një kombinim linear i një numri të njohur të mostrave të mëparshme .

LP (parashikimi linear) është një metodë e modelimit të fjalës që përdor gjërësisht modelin all-pol .Përhapja e kësaj metode buron nga aftësia e saj për të vlerësuar mbështjellësen spektrale të sinjalit të zërit dhe për ta përfaqësuar këtë informacion nga një numër i vogël parametrash.Me modelimin e mbështjellëses spektrale LP kap shenjat më themelore akustike të fjalës ,me origjinë nga dy pjesë të mëdha të mekanizmit të prodhimit të zërit të njeriut ,rrjedhja glottal(i cili është burimi fiziologjik pas cdo structure të mbështjellëses spektrale) dhe trakti vocal(i cili është shkaku i rezonancës locale të mbështjellëses spektrale ,the formant).Përveç aftësisë së saj për të shprehur mbështjellësen spektrale të fjalës me një grup të ngjeshur të parametrave ,LP është e njohur për të siguruar qëndrueshmërinë e të gjithë poleve të modelit,me kusht që të përdoret kriteri i autokorrelacionit.Përveç kësaj njihet që performanca e LP përkeqësohet në prani të zhurmës prandaj zhvillohen disa metoda të parashikimit linear me një qëndrueshmëri kundër zhurmës .Parashikimi linear i pondëruar

,WLP,përdor peshën e doment kohë të rrënjës katrore të sinjalit të gabimit të parashkuar.WLP ka treguar kohët e fundit që jep një mbështjellëse të përmirësuar spektrale të fjalës së zhurmshme në krahasim me analizën tradicionale LP .Në ndryshim me shumë metoda të tjera të fuqishme të LP filtri i parametrave të WLP , mund të llogaritet pa ndonjë përsëritje të të dhënave.

Rrjeti Artificial neural (ANN) shërben si object për sigurimin e një modelii cili ka aftësinë për të lidhur hyrjet dhe daljet e bashkësive të të dhënave shumë komplekse. Ky model ANN punon shumë mirë për grupe shumë komplekse të të dhënave të cilat janë zakonisht shumë të vështira për t'u parashikuar duke përdorur modelimin matematik (ekuacionet).E rëndësishme është trajnimi i rrjetës e cila ka të bëjë me gjetjen e vlerave optimal të peshave të ndryshme dhe animeve të rrjetit.Normalisht ,përdoren lloje të ndryshme të teknikave për të gjetur vlerat e përshtatshme të peshave dhe animeve të ANN.Në këtë punim ,për trajnimin optimal të rrjetit është përdorur teknika e optimizimit të tufës së grimcave ,PSO.Është propozuar metoda apo përafrimi i trajnimit të rrjetës neural artificial duke përdorur PSO.Është përdorur gjuha e programimit Matlab për një trajtim të tillë ,duke aplikuar një databazë të ndërtuar nga shkarkimi i një lajmi shqip shkarkuar nga interneti.

Kapitulli I

Ky kapitull fillon me një përkufizim të nivelit të lartë të koncepteve të sinjaleve dhe sistemeve. Kjo pasohet nga një hyrje e përgjithshme në perpunimin dixhital te sinjalit (DSP), duke përfshirë edhe një pasqyrë të komponentëve bazë të një sistemi të DSP dhe funksionalitetin e tyre. Disa shembuj praktike të aplikimeve DSP jetës reale janë diskutuar shkurtimisht. Kapitulli përfundon me një prezantim të skemat e kursit.

1.1 Konceptet e sinjalit dhe sistemet

Sinjal :

- Një sinjal mund të përkufizohet gjerësisht si ndonje sasi që shkon si një funksion i kohës dhe / ose hapësirës dhe ka aftësinë për të përcjellë informacion.
- Sinjalet I gjejme kudo në shkencë dhe inxhinieri. Shembujt përfshijnë:
    - Sinjale elektrike: rrymat dhe tensionet në qarqet AC, komunikim, sinjale radio, audio dhe sinjalet video.
    - Sinjalet mekanike: zanore apo presion valët, dridhjet në një strukturë, tërmetet.
    - Sinjalet Biomjekësi: elektro-encephalogram, mushkërive dhe monitorimit të zemrës, X-ray dhe llojet tjera të imazheve.
    - Finance: Ndryshime kohë të një vlerë të aksioneve ose një indeks të tregut.
- Njekohesisht, çdo seri të matjeve të një sasie fizike mund të konsiderohet si një sinjal (matjet e temperaturës për shembull).

Karakterizimi i sinjalit :

- Paraqitja më e përshtatshme matematikore të një sinjal është nëpërmjet konceptit të një funksioni, x (t).
  Në këtë shënim:

    - x përfaqëson variablin e varur (psh, tension, presion, etj)

- t paraqet variablën e pavarur (psh, koha, hapësirë, etj).

- Në varësi nga natyra e variablave të pavarur dhe të varur, lloje të ndryshme të sinjaleve mund të identifikohen :

- Sinjal Analog: $t\epsilon\mathbb{R} \rightarrow x_a(t)\epsilon\mathbb{R}\ ose\ \mathbb{C}$

  Kur t tregon kohën, ne gjithashtu i referohen një sinjal të tillë si një sinjal të vazhdueshëm në kohë.

- Sinjali diskret: $n\epsilon\mathbb{Z} \rightarrow x[n]\epsilon\mathbb{R} ose\ \mathbb{C}$

  Kur Indeksi n paraqet vlerat vazhdueshme të kohës, ne i referohemi një sinjal të tillë si diskret ne kohë.

- Sinjali Digital: $n\epsilon\mathbb{Z} \rightarrow x[n]\epsilon A$

  ku $A = \{a_1, a_2, \dots, a_L\}$ përfaqëson një grup të caktuar të L niveleve te sinjalit.

  - Sinjali me shume –kanale : $x(t) = (x_1(t), \dots, x_N(t))$

  - Sinjali multi-permasor : $x(t_1, \dots, t_N)$;

- Dallimet mund të bëhet në nivelin e modelit, për shembull: nëse $x[n]$ është konsideruar të jetë determinist apo i rastit në natyrë.

Shembull 1.1 : Sinjali i fjalës

Një sinjal i fjalës përbëhet nga ndryshimet në presionit e ajrit sin jë funksion i kohës ,kështu që ai paraqet në thelb një sinjal të vazhdueshëm në kohë$x(t)$. Ai mund të regjistrohet me anë tën jë mikrofoni që përkthen variacionet e presionit local (shtypja e presionit të ajrit) në një sinjal të tensionit .Nëse dikush dëshiron që ta përpunojë sinjalin këtë sinjal në kompjuter ,sinjali duhet të jetë diskretizuar në kohë në mënyrë që kompjuteri ta ketë më të lehtë të bëjë përpunimin diskret në kohë , dhe gjithashtu kuantizuar në mënyrë që të akomodohet paraqitja e saktësisë.

Siç e dimë nga teorema e mostrave, sinjal i vazhdueshëm në kohë mund të rindërtohet nga mostrat e marra të saj me një normë mostrave të paktën dy herë më të lartë se komponenti I

frekuences në sinjal. Sinjalet e fjalës shfaqin energji deri , 10 kHz. Megjithatë, shumica e kuptueshmëri është përcjellë në një gjerësi më pak se 4 kHz. Në telefoninë dixhital, sinjalet e fjalës filtrohen (me një filtër anti-aliasing që heq energjinë mbi 4 kHz), mostrohen në 8 kHz dhe paraqiten me 256 nivele diskrete (jo uniformlyspaced). Banda e gjere e fjales (shpesh të quajtur cilësi komenti) do të kërkojë marrjen e mostrave në një normë më të lartë, shpesh 16 kHz.

1.2 Përpunimi dixhital i sinjalit

Perpunimi dixhital i sinjalit është një nga teknologjitë më të fuqishme që do të formojnë shkencën dhe inxhinierinë në shekullin e njëzet e një. Ndryshime revolucionare tashmë janë bërë në një gamë të gjerë fushash: komunikimit, Imazhe mjekësore, radar dhe hidrolokator, besnikëri të lartë riprodhimin e muzikës dhe kërkimet e naftës, për të përmendur vetëm disa. Secila nga këto fusha ka zhvilluar një teknologji të thellë DSP, me algoritme e veta, matematikën, dhe teknika të specializuara. Arsimi i DSP përfshin dy detyra: të mësuarit e koncepteve të përgjithshme që vlejnë për këtë fushë, si një e tërë, dhe të nxënit e teknikave të specializuara për zonën tuaj të veçantë të interesit.

Njohja e fjalës dhe DSP

Njohja e të folurit është shembulli klasik i gjërave që truri i njeriut bën mire dhe kompjuteri ben keq. Kompjutera dixhitale mund të ruajnë dhe të kujtojnë sasi të madhe të të dhënave, të kryer llogaritjet matematikore me shpejtësi që flakëron, dhe të bëjë detyrat e përsëritura pa u mërzitur ose joefikase. Për fat të keq,ne ditet e sotme kompjuterat performojne shumë dobët kur përballen me të dhëna të papërpunuara shqisore. Mësimi ne një kompjuter të ju dërgoj një faturë mujore elektrike është e lehtë. Mësimi ne të njëjtin kompjuter për të kuptuar zërin tuaj është një ndërmarrje e madhe.

Përpunimi dixhital i sinjalit në përgjithësi përafron problemin e njohjes së zërit në dy hapa:

a) *nxjerrjen etipareve* e ndjekur nga b)*përputhshmeria e tipareve*. Çdo fjalë në sinjal hyrje audio është e izoluar dhe më pas analizohen për të identifikuar llojin e eksitimit dhe frekuencat rezonante. Këto parametra pastaj janë krahasuar me shembujt e mëparshëm të fjalëve të folura për të identifikuar perputhshmerine më të afërt. Shpesh, këto sisteme janë të kufizuara në vetëm disa qindra fjalë; mund të pranojë vetëm të folurit me pauza të ndryshme mes fjalëve; dhe duhet të ri-trajnohet për secilin altoparlant individual. Ndërsa kjo është e

mjaftueshme për shumë aplikimet komerciale, këto kufizime janë modest, kur krahasohet me aftësitë e dëgjimit të njeriut.

1.2.1 Mostrimi

Së pari fjalimi mostrohet apo kampjonimi në një frekuencë të përshtatshme për të kapur të gjithë komponentët e nevojshëm të frekuencave të rëndësishme për përpunimin dhe njohjen e fjalës .Mostrimi në përpunimin e fjalës është reduktimi i sinjalit të vazhdueshëm në një sinjal diskret të fjalës .Një mostër apo kampion është një vlerë apo një bashkësi vlerash në një moment të caktuar të kohës apo hapësirës.Në njohjen e të folurit ,normat e përbashkëta për marrjen e mostrave janë 8 kHz deri në 16 kHz .Për të matur me saktësi një valë është nevojshme që të ketë të paktën 2 mostra në cdo cikël ,njeri për të matur pjesën positive të valës dhe tjetri për të matur atë negative , dmth sipas teoremës Nyquist ,marrja e mostrave të frekuencave duhet të jetë të paktën 2-fishi i bandwithit të sinjalit me kohë të vazhdueshme për të shmangur (aliasing)ndryshimet.Për transmetimin e zërit frekuenca e mostrimit zakonisht zgjidhet 10 kHz, por shumë e zakonshme është dhe zgjedhja e frekuences 8 kHz.Kjo zgjedhje ndodh për shkak se për të gjithë folësit e gjithë energjia e rëndesishme e të folurit përmbahet në frekuencat nën 4 kHz (edhe pse disa grad he femije e shkelin kete supozim).Pra shumica e informacionit në fjalën e njeriut është në frekuencat nën 10 kHz ,pra për një normë mostrimi 20 kHz do të jetë e nevojshme për saktësi të plotë.Procesi i mostrimit kalon ne keto hapa :

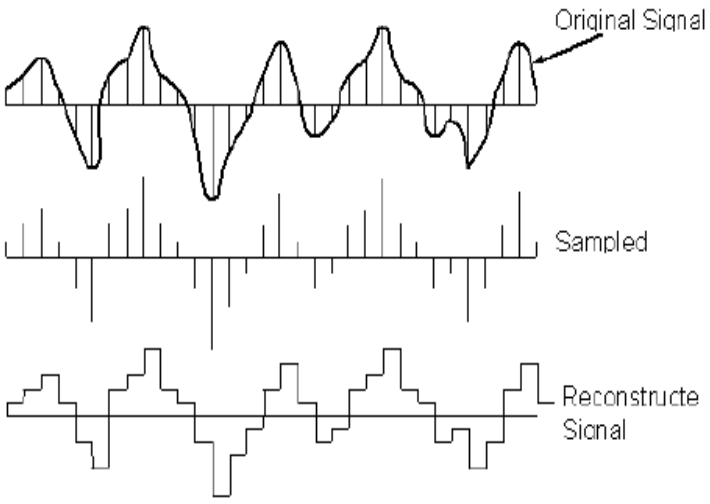


Figura 1.1 Mostrimi i sinjalit

a)Fillimisht sinjali analog *s(t)* është konvertuar në një formë diskrete të kohës *x(t)* duke mostruar vlerat e tij(në një kuptim me specific të fjalës)në interval periodike me gjatësi t_s,periusha e mostrimit .(figura 2.1) Norma e mostrimit përcaktohet si : $f_s = \frac{1}{t_s}$Indeksi *n* isinjalit të mostruar lidhet me kohën $t=nt_s$ i tillë që$x(n)=s(nt_s)$. Pastaj mostrat me vlera të rrumbullakosura japin njëvlerë diskrete numrash binar,që quhet zakonisht *mostër*.Për shumicën e kërkesave audio komercial është përdorur një format ,për kërkuesit, megjithatë një format prej 32 bit pikash notuese me vlerë që variojne nga -1.0 ne 1.0 janë të preferuar.Ky hap është quajtur *kuantizim.*

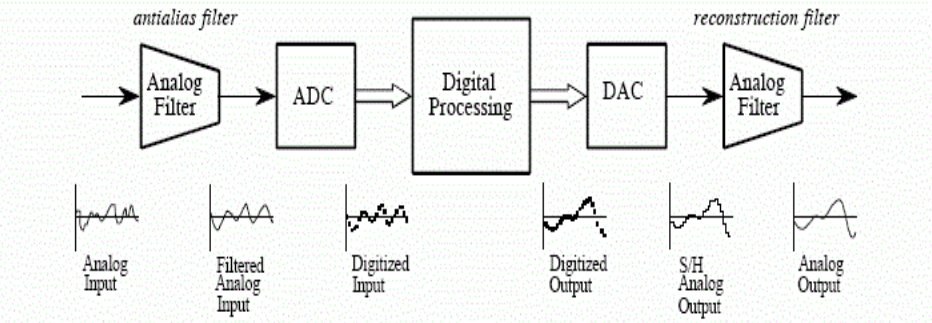


Fig 1.2 .Konvertimi I nje sinjali nga analog ne diskret ne lidhje me kohen (*kampjonimi*)

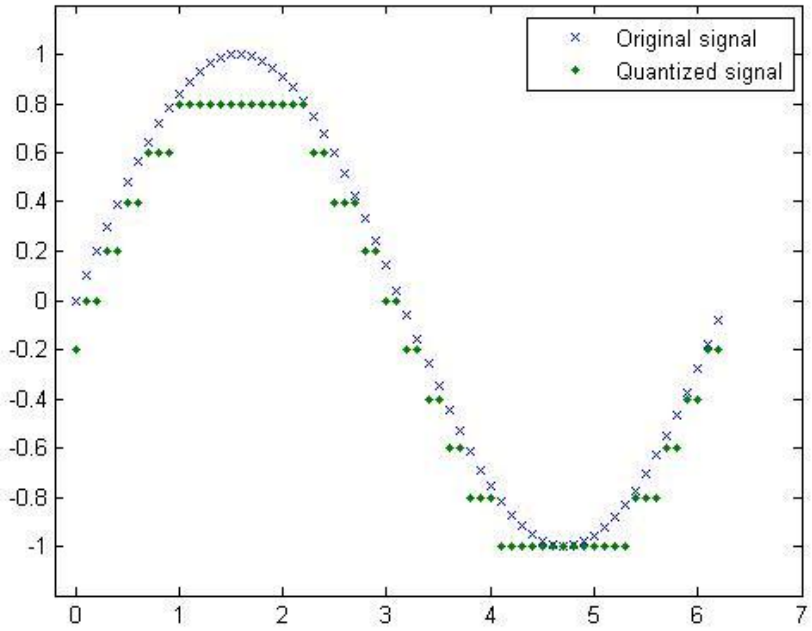


Fig 1.3 . Konvertimi nga sinjal diskret ne kohe ne nje sinjal dixhital (*kuantizimi*)

Rezultati i transformimit analog në dixhital (figura 1.3) është një kurbë me hapa diskret në cdo mostër.Theksojmë që në shumë teori të përpunimit dixhital të sinjalit ,mostrat në përgjithesi konsiderohen me vlera të vazhdueshme.

Ka burime të ndryshme të gabimit në konvertimin A/D.

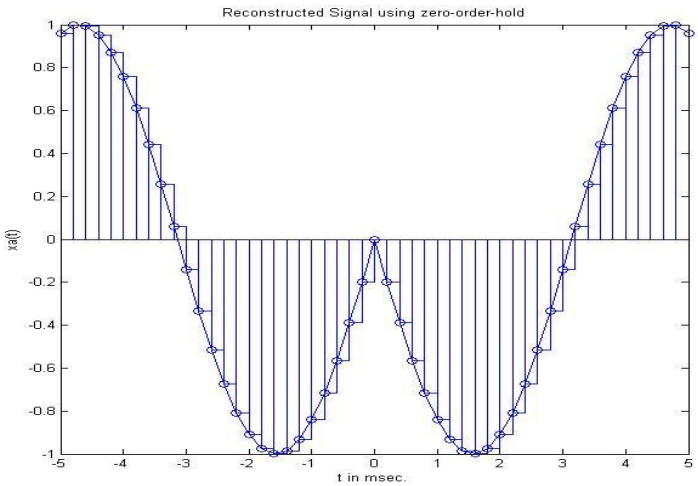


Fig 1.4 . Rezultatet e konvertimit A/D

Aliasing , përcakton keqidentifikimin e një frekuence të sinjalit ,duke përfshirë shtrembërimin apo gabimin .Sinjalet analoge zakonisht kalojnë nëpër filtrin low-pass për të hequr shumicën apo të gjithë komponentet mbi frekuencat Nyquist për të shmangur *aliasing.*

Frekuenca e mostrimit është 10 herë më e vogël se frekuenca sinusoide $f_h = \frac{f_s}{2}$ .Anasjelltas themi qe 10 mostra jane marre nga cdo periudhe e sinjalit dhe kjo eshte mese e mjaftueshme për të ndërtuar sinusoidën e vazhdueshme në kohë.Për ta bërë atë me konkret,imagjinoni sikur sinusoida ka një frekuencë prej 1 Hz dhe norma e mostrimit është 10 Hz(për sinjalet audio nuk është një vlerë praktike).Të shohim se cfarë ndodh kur ne përpiqemi të mostrojmë një sinusoid 8Hz në një normë mostrimi 10 Hz.Sinusoida e frekuencave të larta prej 8Hz prodhon një sekuenë të caktuar të mostrave si zakonisht,megjithatë ekziston një sinusoid e frekuencave 2Hz i cili do të cojë në të njëjtën sekuencë të saktë të mostrave dhe kjo është një sinusoide e frekuencave të ulta e cila .Rregulli i përgjithshëm është : Sa herë që ne mostrojmë

një sinusoidë të një frekuence f mbi gjysmën e normës së mostrimit por nën normën e mostrimit ($f_s > f > \frac{f_s}{2}$) atëhere *alias* i saj do të shfaqet në një frekuencë$f' = f_s - f$ ose $f' = \frac{f_s}{2} - (f - {f_s}/{2})$,ku forma e dytë e bën të qartë se ne duhet të pasqyrojmë tepricën e frekuencës së sinjaleve që gjenden mbi frekuencën Nyquist në frekuencën Nyquist për të marrë*alias* e saj.

1.2.2 Kuantizimi

Pas marrjes së mostrave kemi një sekuencë të numrave të cilat teorikisht mund të marrin në ndonjë vlerë në një gamë të vazhdueshme të vlearave.Meqë kjo gamë është e vazhdueshme ka pafundësi vlerash të mundshme për cdo numër .Në mënyrë që të jetë në gjendje që të paraqesi cdo numër nga një varg i tillë i vazhdueshëm , do të na duhet një numër i pafund i shifrave dicka që ne nuk e kemi.Në vend të kësaj ne duhet të paraqesim numrat tanë me një numër të caktuar të shifrave dmth : pas diskretizimit të variablit kohë ,ne duhet të diskretizojmë variablin amplitudë .Ky diskretizim i vlerave të amplitudës quhet *kuantizim* .Me rastin e kuantizimit, një pjesë e sasisë së informacionit humbet. Minimizimi i kësaj humbjeje bëhet duke përdorur më shumë nivele të kuantizimit. Por, numri i madh i niveleve të kuantizimit do të kërkojë më shumë bita për kodim. Numri i bitave që sot përdoren në kodimin e mostrave të të folurit është 16, prandaj numri i niveleve duhet të jetë 2^{16} (65 536).

Kuantizimi i gabimit shpesh quhet *kuantizimi i zhurmës.* Një arsyetim matematik më rigoroz për trajtimin e kuantizimit të gabimit si zhurma e bardhë me shpërndarje të njëtrajtshme është dhënë nga teorema e kuantizimit Widrow ,por ne nuk do ta ndjekim konkretisht .Përcaktojmë raportin e sinjalit –për –zhurmë të një sistemit si raport midis vlerave RMS të sinjalit të dëshiruar dhe vlerave RMS të zhurmës të shprehur në decibel dB:

$$SNR = 20log_{10}\frac{x_{rms}}{e_{rms}}$$

Ku vlera RMS është rrënja katrore e fuqisë (mesatare) së sinjalit.Shprehim fuqinë e sinjalit si P_x dhe fuqinë e gabimit si P_e atëhere :

$$SNR = 10log_{10}\frac{P_x}{P_e}$$

Fuqia e gabimit të kuantizuar jepet nga varianca e variablit shoqërues të rastit dhe del :

$$P_e = \int_{\frac{-q}{2}}^{\frac{q}{2}} \frac{1}{q} e^2 de = \frac{q^2}{12}$$

Merret një sinusoid në njësi amplitude si sinjal i kërkuar,ky sinjal ka një fuqi :

$$P_x = \frac{1}{2}$$

Duke i zëvëndësuar në ekuacionin e mësipërm marim:

$$SNR = 10log_{10}\frac{6}{q^2}$$

Ky ekuacion mund të përdoret për të llogaritur SNR midis një njësie sinusoide të amplitudës dhe gabimit të kuantizuar që është bërë kur ne kuantizuam sinusoidën me një hap kuantizimi *q*.

1.2.3 Fuqia dhe Energjia

Nocioni i energjisë dhe i fuqisë për sinjalet dixhitale është e lidhur me njësinë fizike që i përkasin sinjalit elektrik analog.Ato janë të lidhuar konkretisht me lartësinë e tingullit apo zërit megjithëse në një mënyrë jo parëndësi.

Energjia e sinjali diskret në kohë*x(t)* përcaktohet si:

$$E = \sum_{n=-\infty}^{\infty} |x(t)|^2 dt$$

Fuqia e sinjalit *x(t)* përcaktohet si:

$$P = \lim_{N\to\infty} \frac{1}{2T+1} \sum_{n=-N}^{N} |x(t)|^2 dt$$

Në qoftë se $0 < E < \infty$,atëhere sinjali *x(t)* është quajtur sinjal energjisë.Megjithatë ka sinjale që nuk i kënaqin këto kushte .Për këto sinjale ne marrim në konsideratë fuqinë .Nëse $0 < P < \infty$ atëhere sinjali është quajtur sinjal fuqi.Theksojmë që fuqia për një sinjal të energjisë është zero ($P = 0$) ,dhe energjia për një sinjal të fuqisë është∞.

1.2.4 Transformimi Furie

Transformimi Furie mund të jetë i fuqishëm në kuptimin e sinjaleve të përditshme dhe të probblemeve të gabimeve në sinjale .Edhe pse transformimi Furie është një funksion ikomplikuar matematikor ,ai nuk është një koncept i komplikuar për të kuptuar dhe lidhur me sinjalet që shqyrtohen .Në thelb ai merr një sinjal dhe e thyen atë në valë sinus të amplitudave dhe frekuencave të ndryshme .Le të bëjmë një vështrim në atë kjo shpreh dhe pse është e dobishme.

Kur shikojmë sinjale reale ,zakonisht i shohim ato si një tension që ndryshon me kalimin e kohës .Kjo referohet si domeni kohor.Teorema Furie thotë se cdo formë valore në domenin kohor mund të paraqitet nga shuma e pondëruar e sinusit dhe kosinusit .

Pas konvertimit A/D ,sinjali $x(n)$ paraqitet në domenin kohë si një sekuencë numrash .Sipas se matematikanit Furie ,çdo funksion ,edhe ato jo periodic ,mund të shprehen sin jë shumë sinusoidesh(periodike) që nuk mund të zbërthehen më tej .Në domenin kompleks ,një sinusoide mund të shprehet si $e^{j\omega n} = cos\omega n + jsin\omega n$ me frekuencë këndore $\omega = \frac{2\pi}{N}k$,për *k* frekuencë diskrete dhe *N* banda frekuencash ,ose lidhur me frekuencat në Herz (Hz) si : $\omega = \frac{2\pi}{f_s}f$. Pra për të zbuluar strukturën e frekuencave ,ai mund të konvertohet në një paraqitje në domenin frekuencë nga transformimi diskret Furie (DFT):

$$X(k) = \sum_{n=0}^{N-1} e^{jnk\frac{2\pi}{N}} x(n)$$

X quhet spektri i frekuencave të sinjalit dixhital *x* me gjatësi *N*.Anasjelltas për të përcaktuar *x(n)* në përdorim transformimin diskret Furie të anasjelltë :

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} e^{-jnk\frac{2\pi}{N}} X(k)$$

Rezultati i transformimit Furie të shprehur në formulën e mësipërme janë*N* numra kompleks *X(k)* të cilat përcaktojnë magnitudën e spektrit *M(k)* dhe fazën e spektrit $\varphi(k)$ për një frekuencë diskrete *k* : $M(k) = |X(k)|$

$$\varphi(k) =< X(k) = arctan\left|\frac{re\ X(k)}{im\ X(k)}\right|$$

Magnituda e spektrit jep intensitetin e një sinusoide në frekuencën *k* ,ndërkohë faza e spektrit jep vonesën e tij në kohë (nuk është i rëndësishëm për njohjen e fjalës).Termi specter Furie ose spectër përdoret për magnitude e spektrit.

Ekziston një përgjithësim i transformimit diskret Furie ,i quajtur **Z-transformimi** ,i cili shpesh thjeshton paraqitjen në domenin frekuencë .Z-transformimi *X(k*) për një sekuencë diskrete *x(n),* ku *z* është një variable kompleks ,përcaktohet në trajtën :

$$X(z) = \sum_{n=-\infty}^{+\infty} x(n)z^{-n}$$

Në qoftë se *z* zëvëndësohet me $e^{j\omega k}$ kemi të bëjmë me transformimin Furie.

Kur llogaritet direkt transformimi diskret Furie kemi të bëjmë me një kompleksitet kompjuterik .Duke përfituar nga periodiciteti i sinusoids dhe duke aplikuar ndarjen dhe parimin e ndarjes së problemit në nënprobleme më të vogla ,një llojshmëri e algoritmeve efikas për kompleksitetin janë zhvilluar të cilat kolektivisht njihen si *transformimi i shpejtë Furie (FFT).*

Në përgjithësi këto algoritme kërkojnë që ten jë fuqi e dyshit .Si transformimi diskret Furie dhe inversi i tij ndryshojnë vetëm në shenjën e eksponentit dhe një faktor shkalle ,të njejtat parime të çojnë në transformimin invers Furie (IFFT) .

1.2.5.Dritarizimi

Deri më tani ne e kemi konsideruar sinjalin sin jë i tërë ,dhe në qoftë se ai është me gjatësi të pafundme.Për aplikime praktike ,në mënyrë që të mos i humbasim të gjithë informacionin kur kalojmë nga domeni kohë në domenin frekuencë ,seksione të vogla të sinjalit prej rreth 10-40 ms trajtohen në të njëjtën kohë.Këto seksione të vogla quhen *dritare* ose *korniza (frame)* .Për të shmangur futjen e objekteve në domenin frekuencë ,shkatuar nga ndërprerje në kufijte e dritares ,sinjali shumëzohet për herë të parë nga një funksion dritare ,i cili parashikon një zbehje të butë dhe venitje nga dritarja.

$$x_\omega(n) = x(n)\omega(n)$$

Sinjali i dritarizuar $x_\omega(n)$ konvertohet nga transformimi Furie .Dritaret zakonisht janë të vendosura në mënyrë që ato të mbivendosen apo përputhen (overlap) (figura 2.4).

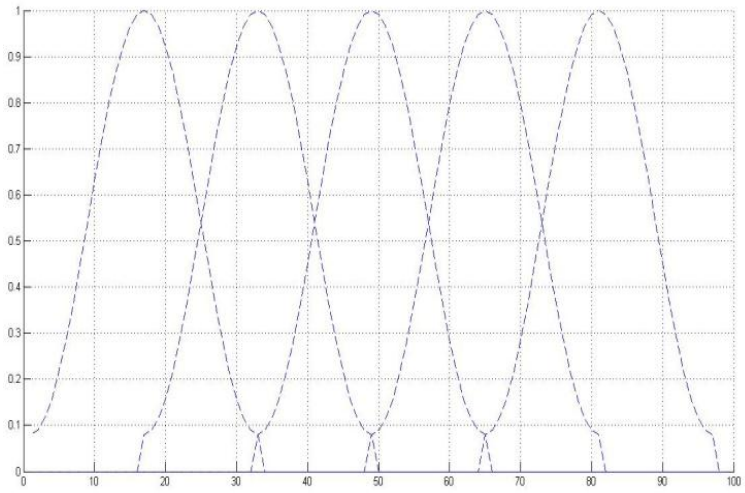


Fig 1.5 . Mbivendosja dritares Hamming(dritarizimi)

Në vazhdim tregohen formulat e disa nga dritaret më të përdorura së bashku me spektrin e magnitudës të një sinjali të përbërë nga dy sinusoida në frekuenca të ndryshme dhe amplitudë ,shumëzuar me funksionin përkatës dritare.Këto spectra tregojnë shtrembërim apo njollosje që është future nga funksioni dritare.Në mënyrë ideale duhet të ketë vetëm dy maja të mprehta.

Drejtkëndore :

$$\omega(n) = \begin{cases} 1 & per\ \ 0 \le n < N \\ 0 & per\ \ te\ tjerat \end{cases}$$

Barlet ose trekëndore :

$$\omega(n) = \begin{cases} n\dfrac{2}{N-1} & per\ \ 0 \le n < \dfrac{N-1}{2} \\ 2 - n\dfrac{2}{N-1} & per\ \ \dfrac{N-1}{2} < n < N \end{cases}$$

Hamming :

$$\omega(n) = 0.54 - 0.46cos\frac{2\pi n}{N-1}$$

Hanning:

$$\omega(n) = 0.5 - 0.5cos\frac{2\pi n}{N-1}$$

Blackman:

$$\omega(n) = 0.42 - 0.5cos\frac{2\pi n}{N-1} + 0.08cos\frac{4\pi n}{N-1}$$

Gauss:

$$\omega(n) = e^{\frac{(n-\frac{N-1}{2})^2}{-2\sigma^2}} \qquad \text{ku } \sigma = \frac{14.41N}{100}$$

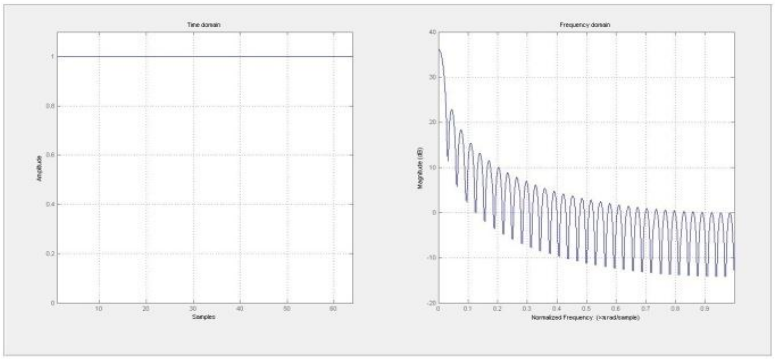


Fig 1.6 Funksioni i dritares *drejtkendeshe* dhe efekti i saj

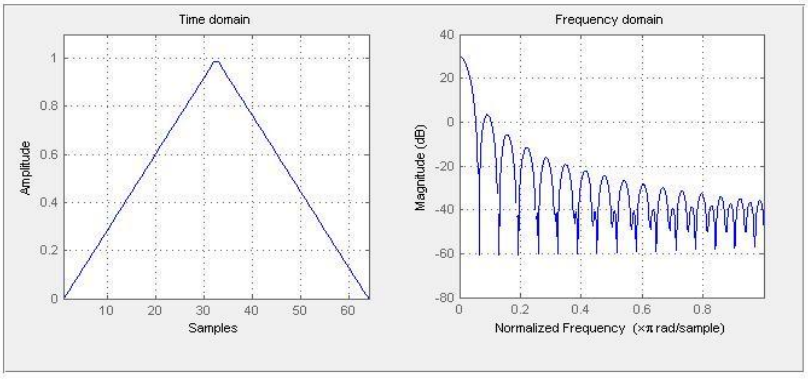


Fig 1.7 Funksioni dritares Barlett dhe efekti I tij

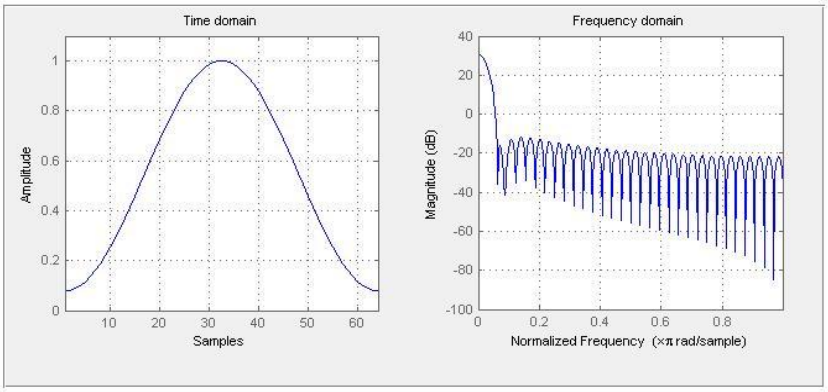


Fig 1.8 Funksioni dritares Hamming dhe efekti i tij

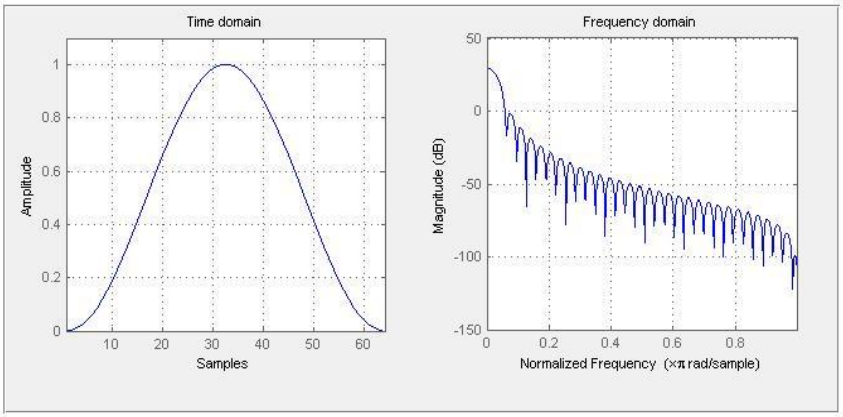


Fig 1.9 . Funksioni dritares Hanning dhe efekti I tij

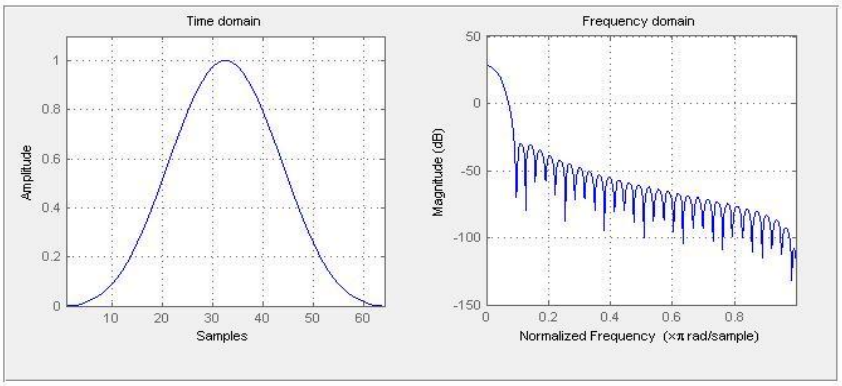


Fig 1.10 Funksioni dritares Blackman dhe efekti I tij

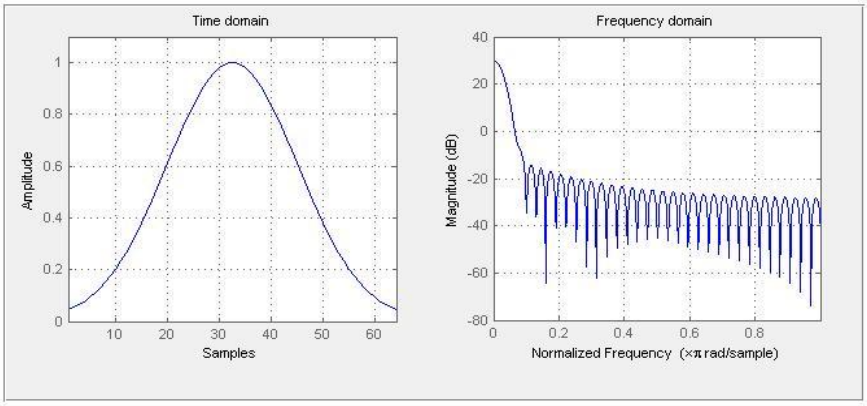


Fig 1.11 Funksioni dritares Gaus dhe efekti I tij

1.2.6 Mbështjellësja spektrale

Shumë gjëra luhaten në universin tonë. Për shembull, të folurit është një rezultat i vibrimit të litarëve vokale të njeriut; yjet dhe planetet ndryshojnë shkëlqimin e tyre se ata rrotullohen në akset e tyre dhe sillen rreth njëri-tjetrin, dhe kështu me radhë. Forma ne domenin kohe e vales nuk është e rëndësishme në këto sinjale; informacioni kyç është në frekuencën, fazën dhe amplitudën e sinusoides përkatese. Transformimi diskret Furie (DFT) është përdorur për nxjerrjen e këtij informacioni.

Një mbeshtjellese spektrale është një kurbë në planin frekuenc-amplitudë, që rrjedhin nga një specter I mashesise Furie . Ai përshkruan një pikë në kohën (një dritare, për të qenë të saktë). Vetitee mëposhtme janë thelbesore për mbeshtjellesen spektrale:

Përshtatja e mbështjellëse

Kurba përshkruan një mbështjellëse te spektrit ,dmth ajo mbeshtillet fort rreth spektrit të magnitudës,duke lidhur majat.

Rregullsia

Një zbutje dhe rregullsi e kurbës është e nevojshme.Kjo do të thotë se mbështjellësja spektrale nuk duhet të luhatet shumë ,ajo duhet të japë një ide të përgjithshme të shpërndarjes së energjisë se sinjalit ndaj frekuencave.

Qëndrueshmëria

Duhet që kurba të jetë e qëndrueshme (në kuptimin matematikor të një funksioni të qendrueshem) dmth ajo nuk ka cepa (kende) (ku derivati i pare kercen).

Mbështjellësja spektrale gjithashtu reflekton klasën e një zanoreje (fonema) në fjalën përkatëse sic mund të shihet në figurën e mëposhtme (2.11).

Pjesët e zërit shfaqen si breze spektrale të ngushta, frekuencat qëndrore të të cilave kanë një marrëdhënie harmonike me frekuencën fundamentale, përderisa te spektret e natyrave tjera është stohastike. Trajta e të gjithë këtyre pikave njihet si *mbështjellsi spektral* (Spectral Envelope), që ekspozon kulmet (majat) dhe gropat (luginat), të njohura ndryshe si *formantet* dhe *antiformantet*.

Ndryshimet e frekuencave qendrore dhe gjërësisë së tyre përcakton timbrin e të folurit përkatës.

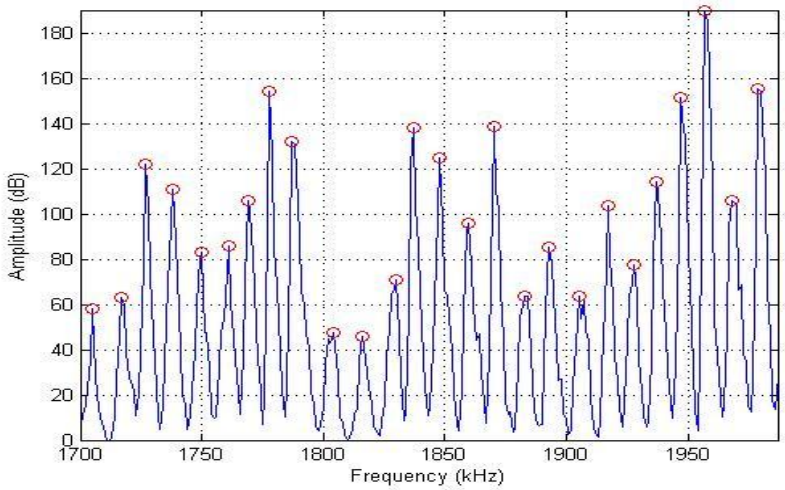


Fig 1.12 Spektri dhe harmonika (ngjyrimi) I nje sinjali

Në fjalën e shprehur ose në zërin e këngëtarit ,mbështjellësja spektrale është mjaft e pavarur nga regjistri i zërit (pitch).Megjithatë në qoftë se ne transpozojmë një zanoreve nga një octavë duke shumëzuar frekuencat e të gjithë pjesëve me 2 dhe kryerjen e një risintetizimi shtesë ,mbështjellësja spektrale do gjithashtu të transpozohet .Ky efekt tingëllon mjaft panatyrshëm .Ky jonatyralitet vjen nga fakti se formantet janë zhvendosur deri në një octavë që korespondon me ngushtimin e traktit vocal për gjysmën e gjatësisë së saj .Natyrisht kjo

nuk është sjellje e natyrshme e traktit vocal.Për të shmangur këtë mbështjellësja spektrale duhet të mbahet konstante ,ndërsa harmonikat (partials) "rrëshqisin" përgjate mbështjellëses me vlerat e tyre të reja.Kjo do të thotë se harmonika e një pjese të transpozuar nuk përcaktohet nga amplituda e pjesës origjinale,por nga vlerat e mbështjellëses spektrale në frekuencat e harmonikave të transpozuara.Në këtë mënyrë vetëm harmonikat janë zhvendosur por mbështjellësja spektrale dhe vendndodhjet e formanteve mbeten të njëjta duke e bërë tingullin e zanores natyral.

1.2.7 Ndikimi i zhurmës

Zhurmë paraqesin pengesat e ndryshme të cilat shfaqen gjatë reproduktimit të zërit të cilat ndikojn në kualitetin e zërit, prandaj zhurmat janë zëra të padëshëruara. Me transmetim do të nënkuptojm pranimin e të folurit dhe muzikës në përgjithësi duke përfshirë edhe ndëgjimin në prani të zhurmave të cilat paraqiten gjatë transmetimit elektroakustik. Ndikimi i zhurmës dhe çdo zërit të padëshiruar, nëse është me intenzitet të lartë, manifestohet me efektin e maskimit. Veshi atëherë fare nuk ndëgjon disa komponente të të sinjalit të dobishëm, e me këtë kualiteti i pranimit bie. Edhe zhurma e intenzitetit më të ulët ndikon në zvoglimin e kualitetit të pranimit edhe atëherë kur është aq e vogël sa që maskohet nën prezemcën e sinjalit të dobishëm. Kjo kryesisht vlen për muzikën. Psh. mjafton që në intervalet e shkurta në mes toneve të dëgjohet një zë i huaj respektivisht zhurma, e që të shkaktoj uljen e kualitetit tek dëgjuesi dhe dekoncentrimin e tij. Ndikimi i zhurmës në kuptushmëri spjegohet në shfaqjen e efektit të maskimit, përgjatë secilës së pari humben komponentet më të dobëta të zërit, në rastin e përgjithshëm zanoret. Prandaj faktori i keqësimit varet jo vetëm nga niveli i zhurmës por varet edhe nga spektri. Nëse është në pyetje zhurma e bardhë ndikimi i sajë është paraqitur në figurën 5.15 e cila vlen për fuqi normale të zërit përafërsisht 70 dB. Nga ky diagram përfundojm se është e mjaftueshme që vlera mesatare e nivelit të zërit te jet vetëm pak decibela mbi nivelin e zhurmës që kuptushmëria të jetë e knaqshme 65% (nën kushtin e transmetimit ideal). Në anën tjetër kuptushmëria bie në 0 atëherë kur niveli i zhurmës është 25 dB e më shum më i lartë se niveli i zërit.

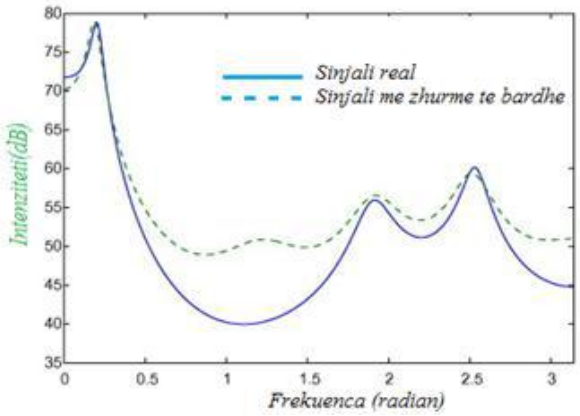


Fig 1.13 : Ndikimi i zhurmës së bardhë në komponentët spectral të sinjalit

Ndikimi i pengesave tjera të cilat nuk kanë karakter të zhurmës, gjithashtu varen nga spektri i tyre që në këtë rast është spektr linjor dhe nga fuqia. Janë llogaritur këto vlera për kuptushmëri por janë shum të komplikuara. Më leht është që problemin ta parashtrojm në formën e kundërt dhe të themi cila vlera të zhurmës nuk guxohet të tejkalohen që kualiteti i transmetimit të jetë në nivel të knaqshëm. Në telekomunikime ekzistojn transmetime me të cilat kjo është rregulluar siq ekzistojn edhe protokola me nivelin e zhurmës së lejuar.

1.3 Metodologjia ,arkitektura e njohjes së fjalës

1.3.1 Karakteristikat e te folurit

Në këtë pjese do të analizohen karakteristikat themelore që janë të rëndësishme për kualitetitn e pranimit të zërave domethënëse për njeriun, dhe të cilat i shërbejnë për përcjelljen e informatave të dobishme (të dëshiruara). Sinjalet më të rëndësishme akustike janë të folurit dhe muzika, por njeriu në jetën e përditshme prodhon, pra edhe pranon, edhe një numër të madh të zhurmave të ndryshme. Pjesa më e madhe e tyre bënë pjesë në grupin e zërave të padëshiruar, por ne këtu nuk do të ndalemi në analizën e tyre sepse me këtë merret një lëmi e veçant e akustikës. Duhet të ceket se analiza e zhurmave nuk bëhet për qëllim të pranimit më të mirë, por me qëllim të nbrojtjes nga veprimi i tyre. Të folurit e prodhon njeriu me organin e të folurit, traktin vokal, i cili përbëhet nga mushkëritë, laringu me kordat e zërit, gjuhëza, zgavra e gojës dhe zgavra e hundës (Fig. 5.1). Diafragma e shtyen ajrin nga mushkëritë nëpër kordat e zërit, duke prodhuar një varg të impulseve të ajrit. Ky varg i impulseve pastaj modulohet me rezonancat e traktit vokal. Rezonancat themelore, të quajtura

formante të zërit, mund të ndryshohen me veprimin e artikulatorëve për të prodhuar zëra të dallueshëm si zanoret.

Kordat e zërit janë të vendosura në fund të laringut në traktin vokal . Procesi i konvertimit të presionit të ajrit nga mushkëritë në vibracione të dëgjueshme quhet artikulim. Kur ajri kalon nëpër kordat elastike të zërit shkakton dridhjen e tyre.

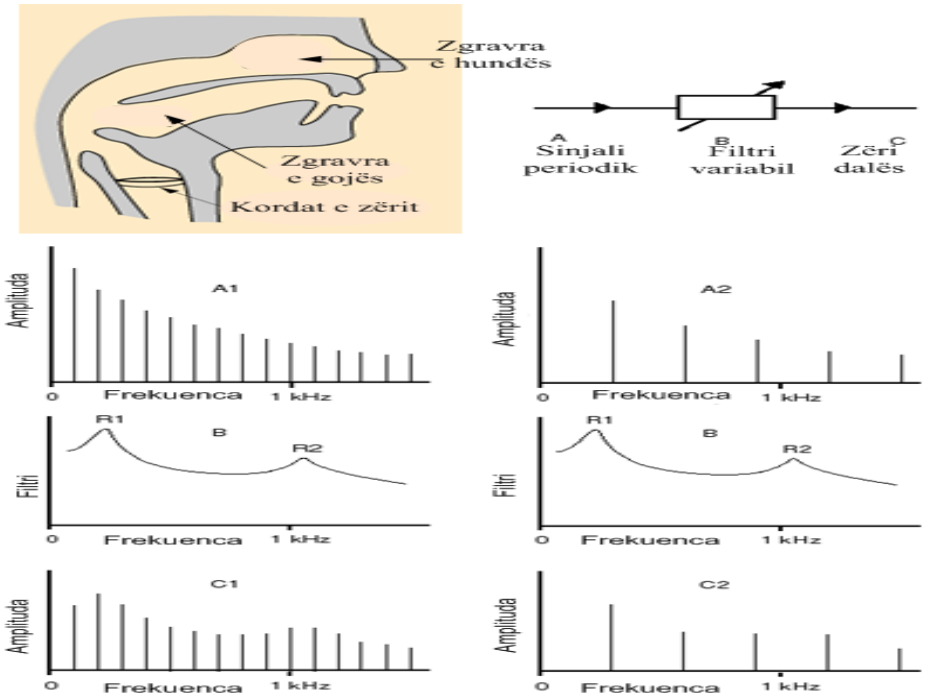


Fig 1.14 : Organi I te folurit

Presioni i ajrit nga mushkëritë i detyron ato të hapen momentalisht, por ajri me shpejtësi të lartë shkakton uljen e presionit sipas efektit të Bernoulli-it, dhe ato pastaj mbyllen menjëherë. Vetë kordat kanë një frekuencë rezonante, e cila e determinon ngjyrën e zërit. Te mashkujtë frekuenca themelore mesatare e të folurit është rreth 125 Hz, te femrat është 210 Hz, ndërsa te fëmijtë zëri mesatar është mbi 300 Hz.

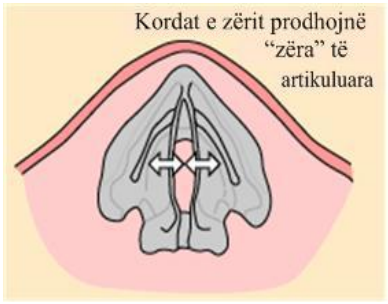


Fig 1.15 : Trakti Vokal

Për prodhimin e zërave të artikuluar, mekanizmi i zërit duhet t'i kontrollojë rezonancat e traktit vokal. Nëse trakti vokal trajtohet si një rezonator , atëherë mund të shifet se pozita e gjuhës, sipërfaqja e hapjes së gojës dhe çfardo ndryshimi që ndikon në vëllimin e zbazëtirës do ta ndryshoj rezonancën.

Zërat fonetikisht ndahen në zanore dhe bashktingllore. Procesi i përshkruar i prodhomit të zërit i përgjigjet zanoreve, sepse prodhimi i bashkktinglloreve ndryshon., te bashktinglloret, në vend të gjeneratorit të impulseve periodike në korda kemi një gjenerator të zhurmës diku në traktin vokal, zakonisht në vend të gjuhës ose buzëve, ku me ngushtimin e traktit vokal ngacmohet rryma e ajrit.

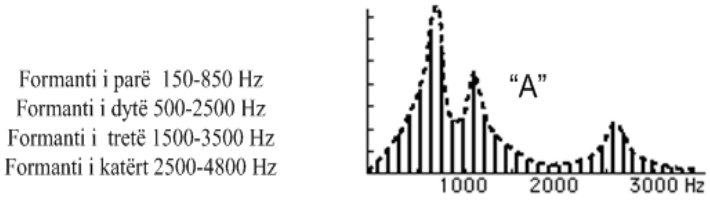


Fig 1.16 : Spektri I zanores A

Zanoret kanë spektër linjor dhe në atë spektër dallohen grupe të harmonikëve të theksuar (Fig. 5.5) të cilët quhen *formante*. Termi formant pra i referohet maksimumeve në spektrin harmonik të tingullit të përbërë i cili paraqitet për shkak të një llojë rezonance të burimit. Për shkak të origjinës së tyre rezonante, ato tentojnë të mbesin të njejta edhe kur të ndryshojë frekuenca themelore (Fig. 5.6). Formantet në tingujt e zërit të njeriut janë rëndsishëm sepse janë komponente esenciale në kuptueshmërinë e të folurit.

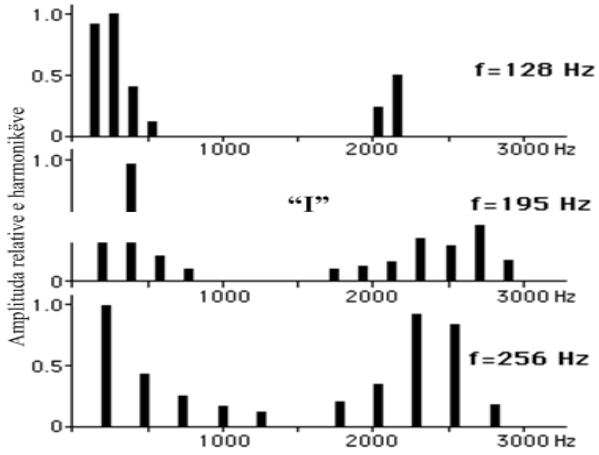


Fig 1.17 : Spektri I zanores I i shqiptuar me frekuence te ndryshme

Paramisht, te çdo zanore janë të theksuara katër formante, por më të rendësishmit për dallimin e zanorës janë dy formantet e para, në bazë të pozitës së të cilave ndryshon spektri i çdo zanoreje .

Zanoret	Brezi frekuencor i formanteve (Hz)
U	200 deri 400
O	400 deri 800
A	700 deri 1200
E	400 deri 700 dhe 1800 deri 2500
I	200 deri 400 dhe 2200 deri 3200

Tabela 1.1 Formantet me te fuqishme dhe me te rendesishme te disa zanoreve.

Pozita e formanteve në shkallën frekuencore varet vetëm nga formësimi i traktit vokal, pra është e qartë se nuk varet nga frekuenca themelore e zërit, e as nga ajo se a është në pyetje zë mashkullor apo femror. Sa i përket frekuencës themelore situata është më ndryshe: vlera mesatare e saj gjat të folurit për zë mashkullor ka vlerën rreth 125 Hz për zë femror rreth 250 Hz ndërsa për atë fëmijëror rreth 300 Hz. Veshi e dallon ngjyrën e zërit duke u bazuar në lokacionin e ngacmimit maksimal përgjatë membranës bazilare të veshit të brendshëm, prandaj membrana bazilare vepron si analizator i spektrit i cili mund të detektoj dallimet në përmbajtjen harmonike me anë të ngacmimeve të ndryshme në vende të ndryshme të membranës bazilare. Pasi që zanoret ndryshojnë para së gjithash në përmbajtjen harmonike të tyre, ky është pra mekanizmi me të cilin veshi i dallon ato .

Bashkëtingëlloret kan spektër kontinual , dhe në të vërehen intervale të theksuara që korrespondojnë me formantet te zanoret. Bashkëtingëlloret megjithatë janë më shumë e me këtë edhe më të rëndësishme për njohjen e rrokjeve dhe fjalëve. Prandaj mund të themi se ato 'bartin' (janë përgjegjëse) për kuptueshmërinë e të folurit. Te bashktinglloret gjithashtu mund të vërehet diçka që i përgjigjet formanteve. Këto do të ishin pjesët e theksuara të spektrit, i cili përndryshe është kontinual si te çdo zhurmë.

Zanoret dhe bashktinglloret së bashku ndërtojnë rrokje prej të cilave përbëhet të folurit. Kohëzgjatja e rrokjeve varet nga shpejtësia e të folurit. Te shpejtësia normale e të folurit kjo kohë është 1/5 sekondës. Pjesa më e madhe e kësaj kohe bie në zanore, sepse bashktingllorja është njëfarë dukurie kalimtare në mes të dy regjimeve stacionare – dy zanoreve – të cilat

mund të zgjasin edhe shum gjatë p.sh. te këndimi. Pjesën më të madhe të energjisë së sinjalit të të folurit e bartin zanoret, posaqërisht komponentet e zërit në regjionet e formanteve.
Ndërkaq, bashktinglloret janë në numër më të madh, prandaj edhe më të rëndsishëm për njohjen e rrokjeve dhe fjalëve. Prandaj mund të thuhet se zanoret e bartin "fuqinë", ndërsa bashktinglloret "kuptueshmërinë" e të folurit. Edhe të folurit i parë si tërësi, e ka spektrin e vet edhe atë të vazhdueshëm që i përfshinë të gjithë zërat. Duke marrë parasysh ndryshimet e vazhdueshme të spektrit derisa flasim, mund të bëhet fjalë vetëm për vlera mesatare për një periudhë të gjatë kohore. Një spektër i tillë shpesh jepet sipas brezit të caktuar të frekuencës, psh. sipas oktavave, por mund të jepet edhe në trajtë të funksionit spektral d.m.th nga 1Hz siç është paraqitur për zërin mashkullor. (Spektri i zërit femror është shumë i ngjashëm, vetëm se mungon një oktavë e ulët).

Një paraqitje shumë domethënëse për të forlurit dhe analizën e tij është edhe spektrogrami. Spektrogrami paraqet pjesë të të folurit në raportin kohë-frekuencë.Spektogrami paraqitet nga nje shkalle ngjyrash si ne figuren e meposhtme ose nga nje shkalle gri .Ne figuren e meposhtme pjeset blu te errata paraqesin pjeset me energji te ulet te sinjalit dhe pjeset me ngjyre te kuqe te ndritshme paraqesin pjeset me energji te larte .

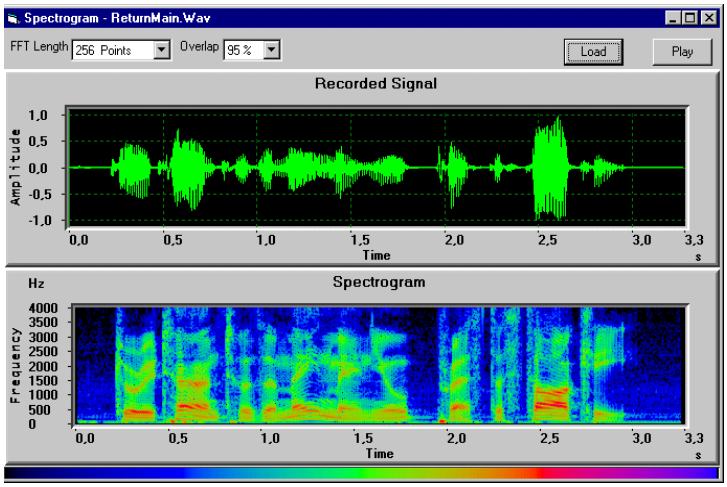


Figura 1.18. Spektogrami i paraqitjes së një sinjali

1.4Paraqitja parametrike e analizës spektrale

Për të zbuluar cilësitë e fjalës ,sinjali duhet të ketë një paraqitje të shkurtër në kohë ,që do të thotë nga vetë natyra sinjali është jostacionar por duke e ndarë apo zbërthyer në sekuenca të shkurtra në kohë atëhere në secilën prej tyre ai konsiderohet stacionar .Gjithashti themi që spektri me kohë të shkurtër I sinjalit është më I përshtatshëm për të shfaqur informacionin fonetik .Në qoftë se do e interpretonim si veprim filtrimi themi që kur fjala prodhohet në kuptimin e një sinjali me kohë të ndryshueshme , karakteristikat e tij mund të paraqiten nga një parametrizim i aktivitetit spectral.Kjo paraqitje e fjalës përdoret në sistemet automatike të njohjes së fjalës ,ku sekuencat e kornizuara konvertohen në një vektor tiparesh që përmbajnë informacionin e përshtatshëm të fjalës.Vektorët kryesor të tipareve janë koefiçientët LP ,bazuar në modelin e prodhimit të fjalës dhe koefiëientët MFC bazuar në modelin e përceptimit të të folurit.

Përpara çdo lloj implementimi faza e pare përfshin përpunimin dixhital të sinjalit e cila konkludon me nxjerrjen e vektorit të karakteristikave apo tipareve ,pastaj faza tjetër pas përpunimit dixhital përfshin përpunimin gjuhësor .

Përpunimi dixhital përfshin hapat nga regjistrimi i sinjalit deri tek konvertimi i sekuencave të kornizuara të sinjalit në një vektor tiparesh të trajtës LPC ose MFCC,që përmbajnë informacionin e pershtatshëm të fjalës . Qëllimi kryesor I nxjerrjes së tipareve është gjetja e një grupi vetish të një fjalimi (shprehje) që ka korrelacion akustik me sinjalin e fjalës dhe nxjerrja e parametrave që llogariten apo vlerësohen me teknikat MFCC dhe LPC.

MFCC përdor spektrin e magnitudes Furie ndërsa LPC përdor spektrin e magnitudes së parashikimit linear .Meqënëse as FFT dhe as LP nuk janë projektuar për të trajtuar kushtet e zhurmës shtesë ,në të cilën një tjetër burim zëri është I pranishëm në mjedisin e regjistrimit apo kanalin e transmetimit .Për këtë arsye janë trajtuar dy modifikime të LP,qe janë WLP dhe SWLP , për të treguar ,në mënyrë të caktuar ,më shumë sjellje të fuqishme në prani të zhurmës shtesë.Këto metoda kanë zëvëndësuar metodën FFT për vlerësimin e spektrit në llogaritjen e MFCC (bankës së filtrave) Përgjigjja impuls e këtyre metodave është përdorur si input në llogaritjen e MFCC.Në kapitullin tjetër trajtohen më gjërë secila metodë.

1.4.1Nxjerrja e tipareve

Për të arritur saktësinë më të lartë në njohjen e të folurit ,zgjedhja e duhur e tipareve nga një sinjal fjalim është problem qëndror .Qëllimi kryesor I nxjerrjes së tipareve është gjetja e një

grupi vetish të shprehjes që ka korrelacion akustik me sinjalin e fjalës,dhe që janë pikërisht parametrat ,që në një farë mënyre mund të vlerësohen dhe llogariten permes përpunimit të formës valore të sinjalit.Parametra të tillë janë quajtur tipare.Ka shumë veti të tipareve që përfshin diskriminimin mes klasave të nënfjalës ,ndryshueshmërine e ulët midis folësve ,degradimin e sinjalit të të folurit për shkak të zhurmës .Metoda të nxjerrjes së tipareve që merren në considerate në këtë punim janë parashikimi linear I koefiçientëve cepstral (LPCC) dhe frekuencat Mel të koefiçientëve cepstral(MFCC).

Shkalla e frekuencave Mel

Çështja kryesore që haste kur modelojmë një system përceptimi apo një fjalim gjenerues është karakteri jolinear i sistemit të dëgjimit të njeriut ,kjo është arsyeja që duhet gjetur një peshore e frekuencave për të modeluar përgjigjen e natyrshme të sistemit të prodhimit të fjalës.Shkalla më e njohur e frekuencave dhe që përdoret në sistemin e njohjes së fjalës është shkalla *Mel* . Kjo shkallë është lineare nen 1 kHz dhe logaritmike mbi 1 kHz,siç tregohet në figurën e mëposhtme.

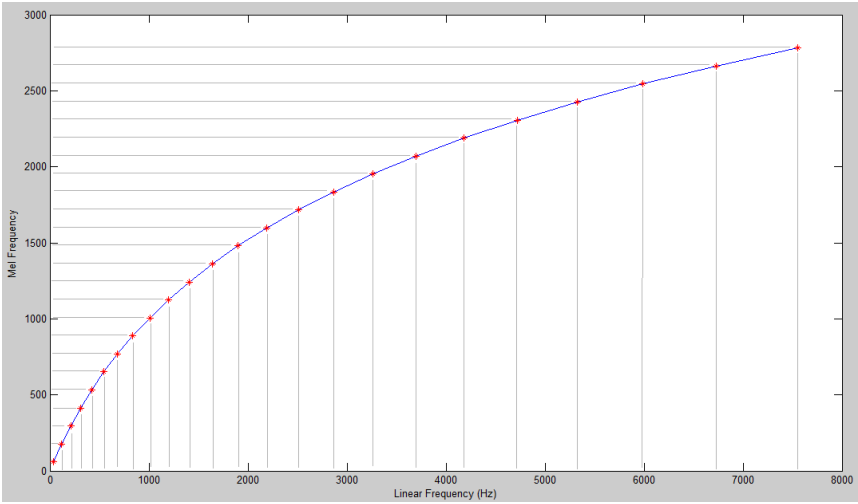


Fig 1.19 Shkalla Mel e frekuencave

Ajo përfaqëson frekuencën themelore(përceptuese)(pitch) të një toni zëri apo regjistri si një funksion i frekuencës së saj akustike .Pra ajo ndërton regjistrat e zërit duke u nisur nga një grup filtrash trekëndësh të mbivendosur .Bashkësia e M dritareve trekëndore të mbivendosura përcaktohet në mënyrë që të ndajë kopetencat e spektrit në shkallën Mel .Për sinjalin e zërit të njeriut që mostrohet në frekuencën 8 kHz (dmth shkalla tipike e mostrave është 8000 mostra për sekondë),një numër prej 24 grupe filtrash konsiderohej i mjaftueshëm .Ky grup I filtrave llogarit spektrin rreth frekuencave qëndrore të cdo bande.f[1],f[2],…f[h] janë përkatësisht

frekuencat më të ulta dhe më të larta njëkohësisht në bankën përkatëse të filtrave e shprehur në Hz dhe $H_1[k],H_2[k],\ldots$ janë dritare trekëndore.Në kapitullin tjetër do sqarohet dhe teknika MFCC dhe ajo LPC ,tani thjesht po japim një skematim të dy teknikave .

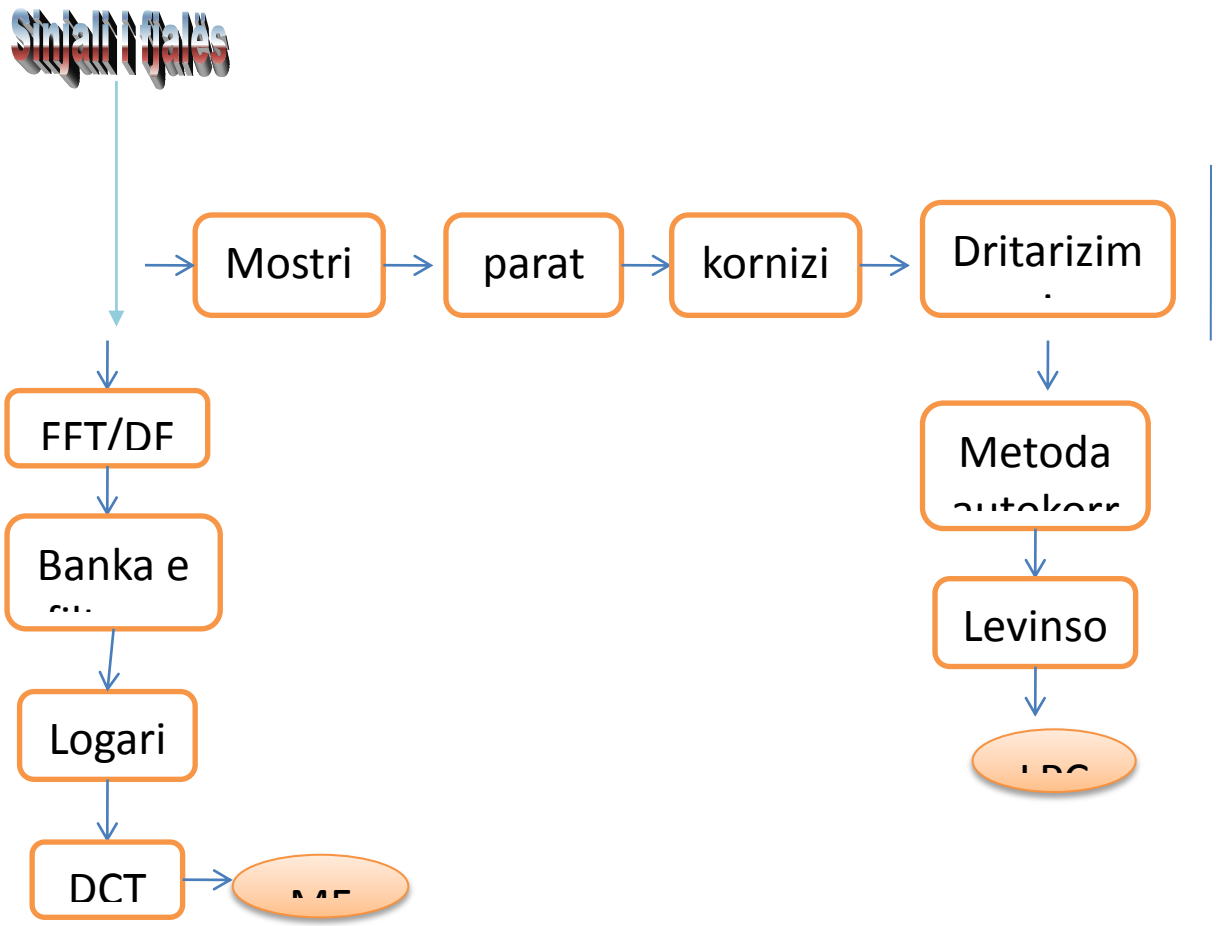


Figura 1.20 . Nxjerrja e tipareve të frekuencave Mel dhe parashikimit linear

Kapitulli II

Paraqitja e sinjalit te fjalës

Edhe pse disa informacione në lidhje me përmbajtjen fonetike mund të nxirren nga ploti i formës valore, kjo nuk përdoret për të ilustruar vetitë e fjalës që janë më të rëndësishme për cilësinë e përgjithshme të tingullit (zërit) ose për përceptimin e detajuar fonetik.

Rëndësia e madhe e rezonancës dhe variacionet e tyre kohore ,përgjegjëse për të mbartur informacionin fonetik ben të nevojshme për të pasur disa mjete për të shfaqur këto karakteristika. Spektri me kohë të shkurtër është më i përshtatshëm për të shfaqur të tilla karakteristika.

Spektri i një sinjali me kohë të shkurtër është madhësia e formësvalore të një transformimi Furie pasi është shumëzuar me një funksion dritare me gjatësi të përshtatshme.Kështu që analiza me kohë të shkurtër bazohet në zbërthimin e sinjalit të fjalës në sekuenca të shkurtra të të folurit ,të quajtura korniza,dhe analiza e secilës prej tyre pavaresisht nga njëra-tjetra.Për të analizuar kornizat ,sjellja (shfaqja periodike ose si-zhurmë) e sinjalit në secilën prej tyre duhet të jetë stacionare(e palevizshme).

Mundte konkludohet që për zgjidhje më të mirëstacionare ,më e përshtatshme është dritarja Hamming pasi ofron rezolucion më të mirë të frekuencave .Në praktikë gjatësitë e dritareve janë 20-30 ms dhe është zgjedhur dritarja Hamming .Kjo zgjedhje është një kopromis midis supozimit të stacionaritetit brënda cdo kornizës dhe rezolutës së frekuencave.

2.1 Modeli i filtrit –burim të prodhimit të fjalës

Nje nga aplikimet kryesore të mbështjellëses spektrale është modeli i prodhimit të të folurit.

Kemi 2 tipe zërash :

Zëri i shprehur

Gjeneruar nga modulimi i shtypjes së ajrit të mushkërive nga hapja dhe mbyllja periodike e kordave vokale

Zëri i pashprehur

Gjenerohet nga një shtrembërim i traktit vocal icili shkakton qarkullim të trazuar të ajrit e cila rezultonnë një zhurmë .

Nga pikëpamja e përpunimit të sinjalit të fjalës ,kjo sjellje mund të modelohet nga një system linear .Burimi është modeluar nga një tren impulsesh për komponentet e fjalës së shprehur ,apo një sinjal i rastit për komponentët e fjalës së pashprehur (shurdhër), duke dhënë një sinjal ngacmim e(n) me transformim Furie E(Z).Burimi actual është formavalore e kordave vokale ose turbullirat e shkaktuara nga një shtrembërim.Spektri i tij është marrë duke filtruar me filtrin burim S(Z).Kështu që efekti i formave të traktit vocal modelohet nga V(Z).Në fund efekti i rrezatimit të buzëve modelohet nga L(Z).

Crregullimet në mushkëri ,kordat apo litaret vocal si dhe ndjenjat e frymëmarrjes janë disa faktorë që ndikojnë në fjalim .Edhe pse prodhimi i fjalës në përgjithësi është një process jo-linear ,për korniza kohore të shkurtra mund të përafrohet në mënyrë të arsyeshme sin jë kaskade e sistemeve lineare ,që përfshinë një sistem burim (burimi i zërit s(n)),një filter all-pol (simulon traktin vocal v(n)) dhe një differentiator (stimulon rrezatimin e buzëve l(n)),domethënë tre pjesët përbërëse të sistemit të prodhimit të fjalës janë :trakti vocal me përgjigje impuls v(n),sinjali ngacmim s(n) që është inputi i traktit vocal dhe rrezatimi i buzëve r(n).

$$h(n) = s(n) * v(n) * l(n)$$

Ku h(n) është sinjali i fjalës ose output i fjalës së shtypur , $*$ shpreh konvolucionin .Duke përdorur z-transformimin ekuacioni i mësipërm mund të shkruhet në domenin frekuence në trajtën:

$$H(z) = S(z)V(z)L(z)$$

Ndërkohë që spektri burim S(Z) dhe rrezatimi i buzëve L(Z) janë në shumicën e rasteve konstante ,të vazhdueshme dhe të njohura a priori ,funksioni i transferimit të traktit vocal V(Z) është pjesa karakteristike për të përcaktuar artikulimin dhe në këtë mënyrë përmbajtjen e fjalës së thënë.Prandaj meriton vëmendje dhe një vështrim nga afër se si mund të modelohet në mënyrë adekuate trakti vocal.

Duke njohur traktin vocal dhe funksionin e tij të transferimit marrim një përkufizim për mbështjellësen spektrale të zërit *: Mbështjellësja spektrale*është funksion transferimi i pjesës së filtrit H(Z) të modelit të burim-filtrit i dhënë me ekuacionin e mësipërm.Kështu për të vlerësuar mbështjellësen spektrale, është e nevojshme që të ndahet H(Z) nga E(Z) ,dhe sinjali fjalim tranformohet me anë të Furies. Meqë sinjali final i fjalës është konvolucion i sinjalit burim me përgjigjen impuls të filtrit H(Z),ky quhet *dekonvolucion.*

Burimi i sistemit të prodhimit të fjalës quhet rrjedhje glotale,e cila luan rol të rëndësishëm në karakteristikat e fjalës.Rrjedhja glotale mund të merret nga sinjali i fjalës duke përdorur një teknikë të quajtur filtrimi i anasjelltë ,ndërkohë që trakti vocal mund të merret nga një filtër all-pol .Në këtë model sic tregohet dhe në fig 2 ,burimi i ngacmimit i ngjan një treni periodic impulsesh duke shkaktuar structurë harmonike si krëhër,ndërsa rezonanca e traktit vocal japin konturin e zhvillimit të zbutur të spektrit pa zhurmë.Një spectër tipik vlerësohet me algoritmin FFT (transformimi fast furie) për një formë valore të fjalës së shprehur .Rezonancat e fort atë traktit vocal quhen *formant* ,dhe ato janë mjetet kryesore për të bërë dallimin midis zanoreve .Në këtë punim metodat e zhvillimit të vlerësimit spectral bazohen në parashikimin linear dhe në modifikime të parashikimit linear duke çuar në disa metoda më të avancuara.

2.2 Vlerësimi i Zhvillimit Spektral

Vlerësimi ka për detyrë të llogarisë mbështjellësen spektrale tën jë sinjali.Për vlerësimin e mbështjellëses spektrale do të përdorim metodat e parashikimit linear dhe modifikime të tij .Këto metoda kanë pikat e tyre të forta dhe të dobta ,në varësi të sinjalit dhe nevojave të përdoruesit .

2.2.1 Kërkesat

Kërkesat për vlerësim janë në thelb përmbushja e vetive të mbështjellëses spektrale të përshkruara më sipër me disa shtesa dhe saktësi.

Përpikmeria

Një arritje e saktë në planin frekuencë-amplitudë e përcaktimit të një sinusoide të pjesshme është njëra nga kërkesat .Më lart kjo u quajt vetia e përshtatshmërisë së mbështjellëses, në të cilën mbështjellësja spektrale është një mbështjellëse e spektrit ,do më thënë ajo rrethon fort magnitudë e spektrit,duke lidhur majat e spektrit.Shkalla e saktësisë përcaktohet nga aftësia përceptuese e veshit të njeriut .Nënormat e ulta të frekuencave ajo mund të dallojë ndryshimet në amplitudë janë aq të vogla sa 1 dB ,dmth ndjeshmëria është më e lartë ,ndërsa në frekuencat e larta ndjeshmëria është pak më e ulët.

Fuqia

Metoda e vlerësimit duhet të jetë e zbatueshme për një gamë të gjërëtë sinjaleve me karakteristika të ndryshme ,nga tingujt me ton të lartë harmonic me hapësirë të gjërë midis pjesëve ,deri tek tingujt e zhurmshëm ose përzierje e tingujve harmonic dhe të zhurmshëm.

Rregullsia

Është e nevojshme një zbutje apo rregullsi e caktuar .Kjo do të thotë se mbështjellësja spektrale nuk duhet të luhatet shumë,por duhet të japë një ide të përgjithshme të shpërndarjes së energjisësë sinjalit përgjate frekuencave.Kjo përkthehet në një kufizim në pjerrësinë(vendin e pjerret) të mbështjellëses (dhënë nga derivati i saj i parë).

2.2.2Analiza e autokorrelacionit

Meqënese sinjali i fjalës është një sinjal jostacionar është e përshtatshme që të përcaktohet një funksion autokorrelacioni me kohë të shkurtër i cili vepron në segmente të shkurtra të sinjalit,dmthënë secili prej sinjaleve të dritarizuar autokorrelohet sipas formulës së mëposhtme :

$$\varphi_l(m) = \frac{1}{N}\sum_{n=0}^{N'-1}[x(n+l)w(n)][x(n+l+m)w(n+m)]$$

Ku $0 \leq m \leq M_0 - 1$, $w(n)$ është një dritare e përshtatshme për analizimin , N është gjatësia e seksionit që bën analizimin , N' është numri i mostrave të sinjalit që përdoret në llogaritjen e $\varphi_l(m)$, M_0 është numri i pikave të autokorrelacionit që janë për tu llogaritur, l është indeksi i fillimit të mostrimit të kornizes.Duke i trajtuar segmented e fjalës së shprehur si periodic për të gjithë analizimin dhe përpunimin ,analiza e autokorrelacionit përdoret kryesisht për llogaritjen e periudhës " pitch" , T_0 për domenin kohë dhe frekuencën "pitch" ose frekuencën themelore F_0 në domenin e frekuencave .Përveç specifikimeve ,termi "pitsh" i referohet frekuencës themelore F_0. "Pitch" është nje atribut i rëndësishëm i fjalës së shprehur .Ai përmban informacionin specific te folësit ,gjithashtu ai është një tipar i nevojshëm për kodimin e fjalës së shprehur ,prandaj vlerësimi i "pitch" është një nga çështjet e rëndësishme në përpunimin e fjalës.

2.2.3 Mbështjellësja spektrale LPC

Zhvillimi spectral është një kurbë në planin frekuencë-amplitudë,që merret apo nxirret nga një spektër i magnitudës Furie. Është një funksion i frekuencës i cili krahason amplitudat e pjesëve të veçanta të spektrit .Është faktori bazë për timbrin e zërit.Zhvillimi spectral ndryshon me kohën dhe mund të ndryshoje frekuencen themelore të zërit.

Metoda standarte për vlerësimin e zhvillimit spectral apo analizimin e spektrit për shumë aplikacione është transformimi diskret Furie i zbatuar nga FFT .Përvec kësaj ,parashikimi linear LP që zakonisht i referohemi si LPC ,është një tjetër metodë e përhapur për modelimin e spektrit me magnitudë me kohë të shkurtër të sinjalit të fjalës.

Në vecanti modeli filter-burim i përdorur në parashikimin linear njihet si modeli parashikimit linear i kodimit. Ai ka dy komponente kryesor : analiza apo kodimi dhe sinteza apo dekodimi(deshifrimi). Pjesa e analizës LPC përfshin shqyrtimin apo ekzaminimin e sinjalit të fjalës dhe ndarjen e tij në segmente apo blloqe .Secili segment shqyrtohet më tej për të gjetur përgjigjen e disa pyetjeve të rëndësishme :

- Është segmenti i shprehur apo i shurdhët
- Cfarë është Pitch i segmentit
- Cfarë parametrash janë të nevojshem per te ndertuar nje filter qe modelon traktin vocal per segmentin actual.

Analiza LPC kryhet zakonisht nga një dërgues që i përgjigjet këtyre pyetjeve dhe zakonisht i transmeton këto pergjigje mbi një marrës .Marrësi kryen sintezen LPC duke përdorur përgjigjet e marra për të ndërtuar një filtër me kusht që burimi i saktë input të jetë në gjendje të riprodhojë me saktësi sinjalin original të fjalës.Në thelb sinteza LPC përpiqet të imitojë prodhimin e të folurit njerëzor .Të gjithë koduesit e zërit kanë tendencë që të modelojnë dy gjera : ngacmimin dhe artikulimin .Ngacmimi është lloji i tingullit që është kaluar në filtrin apo traktin vocal dhe artikulimi është jo normal apo i deformuar.

2.2.3.1 Modeli me Pole-Zero

Parashikimi linear i mostrës n-te , $s[n]$, përcaktohet si kombinimi linear i p mostrave të mëparshme të mostrës $s[n]$. Paraqitja matematike për parashikimin e mostrës $s[n]$shënohet $\hat{s}[n]$ dhe shprehet nga barazimi :

$$\widehat{s[n]} = \sum_{k=1}^{p} a_k s[n-k] + Gu(n)$$

Ku a_k për $1 \leq k \leq p$ shpreh koefiçientët e parshikimit .Ky model që përdoret më shpesh dhe që përbëhet vetëm nga mostrat e mëparshme ose output $s[n]$ i sistemit, mund të përgjithesohet për të përfshirë edhe inputet u[n] të sistemit ,prandaj paraqitja e sinjalit s[n] i përket të ashtequajturës modeli mesatar i lëvizjes autoregressive (ARMA)dhe paraqitet si më poshtë :

$$\widehat{s[n]} = \sum_{k=1}^{p} a_k s[n-k] + G \sum_{l=0}^{q} b_l s[n-l]$$

Ku a_k për $1 \leq k \leq p$,b_l për $1 \leq l \leq q$,dhe G është gain .Eluacioni i mësipërm mund të shprehet në domenin frekuencë në trajtën :

$$H(z) = \frac{S(z)}{U(z)} = G \frac{1 + \sum_{l=1}^{q} b_l z^{-l}}{1 + \sum_{k=1}^{p} a_k z^{-k}}$$

Ku S(z) dhe U(z) janë z-transformimi i s(n) dhe u(n) respektivisht. H(z) është funksioni sistem i sistemit që u modelua i quajtur modeli pol-zero.Rrënjët e numëruesit dhe emëruesit quhen zero dhe pole të sistemit respektivisht .Kjo çon gjithashtu në raste të veçanta të këtij modeli të përgjithshëm .Në qoftë së supozojmë që numëruesi i sistemit të mësipërm është i fiksuar ,i përcaktuar si $b_l = 0$ për $1 \leq l \leq q$,modeli i mësipërm reduktohet në të ashtëquajtërën modeli all-pol (gjithe-pole) ose modeli autoregresiv (AR).

$$H(Z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

Rrjedhimisht ,rasti tjetër mund të përcaktohet ,në qoftë supozojmë që numëruesi është i fiksuar si $a_k = 0$ për $1 \leq k \leq p$,dhe modeli sillet në një model all-zero (vetëm-zero) .Ky quhet modeli me lëvizje mesatare(MA).Në kushtet e modelimit spectral mund të konceptonim se modeli all-pol vlerëson në thelb maksimumin lokal dhe modeli all-zero vlerëson në thelb minimumin local të spektrit. Në modelimin spectral të fjalës është më e rëndësishme që të përshtasim apo përputhim pjesët më energjike të spektrit ,dmth maksimumin local apo formantet e fjalës ,lugina spektrale e të cilës ka vetëm pak energji.Prandaj në përpunimin e fjalës marrin përparësi modelet all-pol.Modelimi all-pol është një metodë shumë e thjeshtë dhe e drejtperdrejtë e cila nuk kërkon një sasi të madhe në

llogaritje ose kujtesës ,veçanërisht në krahasimme modelin all-zero , veçanërisht në krahasim me modelin pole-zero.Për më tepër modeli pole-zero nuk mund të zgjidhet në një formë të mbyllur.

2.2.3.2 Modeli All-Pol

Parashikimi linear është një mjet tradicional i përpunimit të sinjalit për ndërtimin e modeleve statistikore parametrike all-pol.Në kontekstin e përpunimit të fjalës ,parashikimi linear është përdorur për të nxjerrë një paraqitje parametrike për zhvillimin spectral të një sinjali të fjalës.Modeli i prodhuar nga parashikimi linear është një model all-pol i trajtës :

$$H(Z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (1)$$

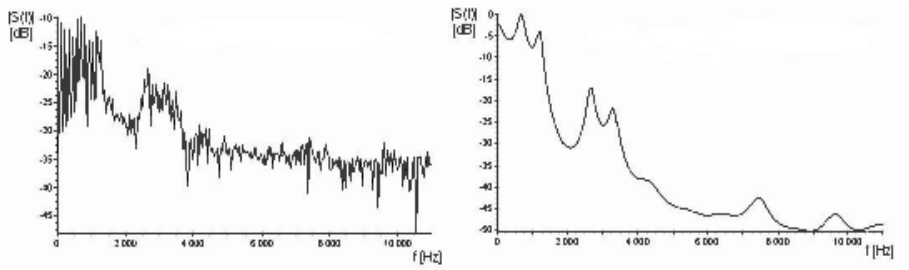


Figura 2.1 : Transformimi Furie dhe funksioni i transferimit të parashikimit linear

Ky modelon mirë një tingull johundor ,ndërsa për tingujt hundor efekti i zerove të përfshira në modelin e detajuar akustik të prodhimit të fjalës duhet modeluar duke përdorur një numër të mjaftueshëm polesh.Rezolucioni spectral i modelit LP varet nga rendi i parametrave të modelit.Meqë rritet rendi i modelit ,modeli LP ndjek spektrin më të ngushtë,por përdorimi i një numri të madh të poleve bën që modeli të fillojë gjurmimin e strukturës harmonike të fjalës,siç mund të shihet në krahasimin e LP dhe modeleve të përgjigjes me shtrembërim me të ulët të variancës minimale (MVDR)(minimum variance distortionless response),të rendit të lartë.Paraqitja parmetrike e zhvillimit spektral ka një përdorim të gjërë.Për shembull një numër relativisht i vogël i parametrave mund të përdoret për të përshkruar mirë tiparet e spikatura të tingullit të analizuar ,parashikimi linear mund të përdoret në aplikimet e kodimit audio,për të koduar ose shifruar një sinjal të fjalës në një kanal me një normë shumë të ulët bitesh.Në kontekstin e këtij punimi ,modelet e zhvillimit spectral të ofruara nga parashikimi

linear ose variantet e tij mund të përdoren për të nxjerrë karakteristikat e fjalës që përformojnë më mirë sidomos në kushtet e zhurmès.

2.3 Parashikimi linear tradicional LP

Sic u tha më lart parasikimi linear është një mjet i fuqishëm modelimi ë fjalës duke prodhuar një model all-pol të sistemit të filtrit ,$V(z)$ i cili është një model i traktit vokal dhe i rezonancës së tij ose *formants.*Sic u përmend me lart inputi i një modeli të tille është një seri pulses ose zhurme të bardhë ,për fjalën e shprehur dhe fjalën e pashprehur respektivisht .

Në modelimin e sinjalit të parashikimit linear një mostër apo kampion $s(n)$ vlerësohet nga një kombinim linear i*p* mostrave të mëparshme .

$$\widehat{s[n]} = \sum_{k=1}^{p} a_k s[n-k]$$

Kështu që duke përdorur LP ,ne mund të vlerësojmë apo llogarisim filtrin e traktit vocal $v(n)$ nga sinjali i fjalës,dhe pastaj të fshijme apo anullojmë ndikimet apo efektet e tij duke rezultuar në burimin e formës së valës.Për të gjetur përgjigjen e traktit vocal ne duhet të zgjidhim një problem minimizimi të katrorit të gabimit ,ku gabimi jepet me anë të formulës.

$$e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k]$$

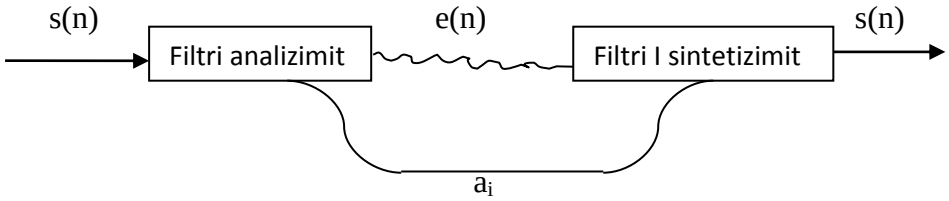


Figura 2.2. Analiza LP

Ku a_k janë koefiçientët e parashikimit linear për këtë sinjal dhe *p* përfaqëson *rendin e modelit* .Minimizimi kryhet mbi një zonë R.Gabimi total jepet nga formula :

$$E = \sum_{R} (s[n] - \sum_{k=1}^{p} a_k s[n-k])^2$$

Nëvarësi nga zona apo rajoni që zgjedhim do të shkojmë në dy teknika të ndryshme të parashikimit linear të cilat së bashku me përmiresimet e tyre do ti paraqesim në kapitujt e tjerë.

Në qoftë se supozojmë se sinjali i fjalës është zero jashtë një intervali $0 \leq n \leq N-1$, atëhere sinjali i gabimit do të jetë e ndryshme nga zero në intervalin $0 \leq n \leq N+p-1$. Ky interval është zona apo qarku R.Meqënëse ne po perpiqemi të parashikojmë mostra të ndryshme nga zero nga mostra zero në fillim të intervalit ,kjo gjë do të rezultoje në gabim të madh në atë pikë .Kjo vlen dhe për vlerat në fund të intervalit ku zero mostrat janë parashikuar nga ato jozero.Për këtë arsye njëdritare konike (p.sh Hamming) përdoret shpesh .Supozimet e lartpërmendura rezultojnë në metodën e autokorrelacionit të parashikimit linear , e cila mund të formulohet si më poshtë :

Duke përdorur dritaren Hamming $w[n]$ me gjatësi N ,ne marrim një segment të fjalës $s_N[n]$ të dritarizuar ,të tillë që$s_N[n] = s[n]w[n]$.Atëhere katrori i gabimit parashikues përcaktohet si më poshtë:

$$E_N = \sum_{n=-\infty}^{\infty} e^2[n] = \sum_{k=-\infty}^{\infty} (s_N[n] - \sum_{k=1}^{p} a_k s_N[n-k])^2 \quad (2.1)$$

Vlerat a_k që minimizojnëE_N gjenden duke marrë derivatet e pjesshme te E_N në lidhje me a_k tëbarabarta me zero , $\frac{\partial E_N}{\partial a_k} = 0$. Kjo ndiqet nga p ekuacione me p variabla të panjohur si më poshtë:

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_N[n-i]s_N[n-k] = \sum_{n=-\infty}^{\infty} s_N[n]s_N[n-i] \quad 1 \leq i \leq p \quad (2.2)$$

Duke vene re qe sinjali I fjales I dritarizuar $s_N[n]$=0 jashte dritares $w[n]$ dhe duke prezantuar funksionin e autokorrelacionit ne trajten :

$$R(i) = \sum_{n=i}^{N-1} s_N[n]s_N[n-i] \qquad 0 \leq i \leq p \qquad (2.3)$$

Atëhere ekuacioni i mësipërm merr trajtën :

$$\sum_{k=1}^{p} R[|i-k|]a_k = R(i) \qquad 1 \leq i \leq p \qquad (2.4)$$

Duke përdorur kuptimin e matricës ,ekuacioni i mësipërm mund të shkruhet në trajtë matricore :

$$\Phi\vec{a} = \vec{r}$$

Ku matrica Φ quhet matrice e autokorrelacionit me elemente te trajtes $\Phi_{i,j} = R(|i-j|)$, ku $1 \leq i,j \leq p$, dhe dy vektoret e tjere jane te trajtes $\vec{a} = [a_1, a_2, \dots, a_p]^T$ dhe $\vec{r} = [R(1), R(2), \dots, R(p)]^T$.

Avantazhi më i madh i metodës së autokorrelacionit është se kjo metodë gjithmonë prodhon një filtër të qëndrueshëm me një ngarkesë të arsyeshme kompjuterike dhe gjithashtu ka algoritme të shpejtë sic është algoritmi rekursiv Levinson-Durbin , për zgjidhjen e sistemit matricor .Efektet e problemeve të analizimit të qarkut apo rajonit R në fillim dhe në fund të saj mund të reduktohen duke përdorur një dritare jo-drejtkendore sic është p.sh dritaria Hamming.

Disavantahet e parashikimit linear LP

-Priret që të mbështjell spektrin sa më fort të jetë e mundur ,dhe në kushte të caktuara zbresin deri në nivelin e zhurmeë se mbetur ne hendekun midis dy pjeseve harmonike.Kjo do të ndodhë kur hapesira midis dy pjeseve është e madhe si në kulmet e larta të tingujve dhe rendi i filtrit është shumë i madh.

2.3.1 Cepstrumi I mbështjellëses spektrale

Cepstrumi është një metode e analizës së të folurit bazuar në një paraqitje spektrale te sinjalit. Për të shpjeguar ide të përgjithshme të metodës se cepstrumit të përdorur për vlerësimin e zhvillimit spektral , dy veti jane te mundshme. Së pari, ne thjesht mund të mendojme për marrjen e mbeshtjelleses spektrale nga një spektër magnitudës Furie nga zbutja e kurbës e saj për të shpetuar nga luhatjet eshpejta. Kjo zgjat deri ne aplikimin e një filtri low-pass për

spektrit, interpretimit si një sinjal, i cili lejon vetëm luhatje të ngadaltë (oshilacione me frekuencë të ulët të kurbës së) .Së dyti, duke kujtuar se një sinjal mund të shihet si konvolucion të një sinjal burim me një filtër, ne duhet të ndajme spektrin burim nga funksioni transferimit i filtrit, e cila është një vlerësim shumë i mirë i mbeshtjelleses spektrale.

Sipas modelit të burim-filtrit tëprodhimit të fjalës një sinjal i fjalës *x(n)* mund të shprehet sin je konvolucion i një burimi ose sinjali nxites *e(n)* dhe përgjigjes impuls te filtrit të traktit vocal *h(n).*

$$x(n) = e(n) * h(n)$$

Në domenin frekuencë ky konvolucion bëhet shumëzim i transformimeve Furie përkatëse.

$$X(\omega) = E(\omega)H(\omega)$$

Duke marrë logaritmin e vlerës absolute të transformimeve Furie ,ekuacioni konvertohet në një shumë :

$$log|X(\omega)| = log|E(\omega)| + log|H(\omega)|$$

Nëse ne aplikojmë një transformimi Furie ne logaritmin e magnitudës së spektrit ne marrim shpërndarjen e frekuencave të luhatjeve apo lëvizjeve në kurbën e spektrit ,e cila quhet *cepstrum.*

$$c = F^{-1}(\log|X(\omega)|) = F^{-1}(log|E(\omega)|) + F^{-1}(log|H(\omega)|)$$

Nën supozimin e arsyeshëm se burimi i spektrit ka vetëm luhatje të shpejta (sinjali ngacmim *e* është i qëndrueshëm ,luhatje të rregullta rreth 100 Hz),kontributi i tij tek *c* do te jetë i përqëndruar në zonat e tij më të larta ,ndërkohë që kontributi i*H* ne rastin kur kemi luhatje të ngadalta të spektrit*X* ,dhe do te jetë i përqëndruar vetëm në pjesën e poshtme të*c*,sic mund të shihet në figurën e mëposhtme .Kështu ndarja e dy komponeteve bëhet e parëndësishme .Vetëm koeficient e parë cepstral*p* ,$c_1, c_2, \ldots c_p$te koeficienteve cepstral c_i mbahen, ku *p* quhet rendi i cepstrumit .Këto përfaqësojne komponentët e frekuencave të ulta ,dmthëne ndryshmet e ngadalta të luhatjeve ,nga zbutja e spektrit*X* për tu bërë një mbështjellëse spektrale.

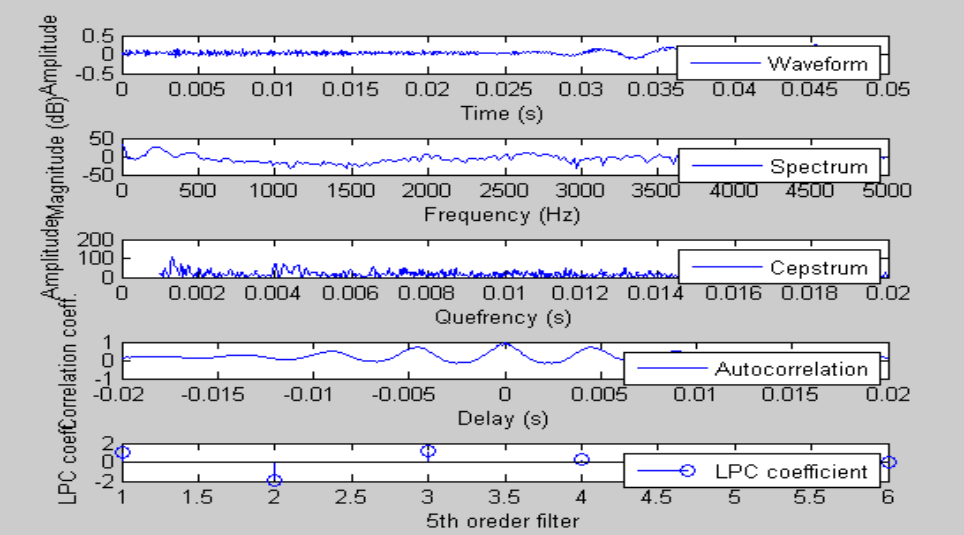


Figura 2.3 Paraqitjet e filtrit linear

Megjithate ,as FFT as LP nuk janë projektuar për të trajtuar kushtet e zhurmës shtesë në të cilën një tjetër burim zëri është i pranishëm në mjedisin e regjistrimit apo kanalin e trasmetimit.Për këtë, arsye në kushtet e zhurmës , janëdiskutuar dy modifikimet e LP për të treguar në mënyra të caktuara,sjelljen me të fuqishme në prani të zhurmës shtesë . WLP dhe SWLP janë një përgjithësim i drejtpërdrejtë i LP që përdorin një funksion të peshuar kohor në mënyrë që të përqëndrohet në zona më pak të zhurmshme të sinjalit.

Metoda SWLP është një modifikim i parashiki,it linear që garanton stabilitetin e modelit all-pol dhe kështu është i përshtatshëm për kodimin dhe sintetizimin e fjalës.

Ndërkohë metoda WLP ,në ndryshim nga LP tradicionale , përdor pesha jo-uniforme të katrorit të mbetjes për tëtheksuar disa zona dhe për të de-theksuar disa te tjera ne aspektin e modelimit të energjisë së gabimit .

Idea e peshës është motivuar nga pikëpamja e llogaritjes se modeleve të parashikimit linear të përmiresuara (WLP dhe SWLP) të cilat janë më të fuqishme kundër zhurmës se sa metoda tradicionale LP .Kjo prespektive është e bazuar në faktin se funksioni peshë i përdorur në këtë rast STE ,mbipeshon ato seksione të formesvalore të të folurit të cilat përbëjnë mostra me amplitudë të madhe .Këto segmente janë më pak të rrezikuar për të shtuar zhurmën e shpërndare uniformisht në krahasim me segmented me vlera me amplitudë të vogël.

2.4 *Parashikimi LinearPonderuesiStabilizuar* (*SWLP*)

Parashikimi linear i pondëruar (WLP) bazohet në aplikmin e një termi që do ta quajmë peshë kohore në funksionin e parashikimit linear konvencional në ekuacionin (3) .Funksioni me peshë kohore mund të përdoret për të drejtuar apo udhëhequr algoritmin e parashikimit linear LP të fokusohet ne modelime të zonave të veçanta kohore të sinjalit input .Në rastin e fjalimit te zhurmshëm ,për shembull,rëndësi i jepet zonave kohore me energji të lartë ,ku efekti relative i zhurmës është më pak i dukshëm.Në praktikë ,ponderimi zbatohet duke zëvëndesuar matricen R të autokorrelacionit te ekuacionit të mësipërm me një matrice të ponderuar të autokorrelacionit .

Parashikimi linear pondërues i stabilizuar (SWLP) i prezantuar nga Magi,është një metodë e modelimit të filtrave all-pol e bazuar në parashikimin linear pondërues (WLP) .WLP përdor peshën apo koeficientin e domenit kohë të sinjalit te katrorit të gabimit parashikues .Kjo peshë kohore thekson mostrat e fjalës të cilat kanë një sinjal me normë të lartë të zhurmës dhe kështu ajo dëshmon qe WLP përmireson vlerësimin e zhvillimit spectral të fjalimit të zhurmshëm në krahasim me analizën tradicionale të parashikimit linear .Për me tepër ne krahasim me metoda të tjera të fuqishme të parashikimit linear ,parametrat e filtrit WLP mund të llogariten pa ndonjë perditësim përsëritës .Një problem i filtrit WLP është se ky filter nuk siguron stabilitetin e tij ,pra nuk është një filtër stabil .Kjo pengese kalohet me ndihmën e filtrit SWLP ,ku koeficienti apo pesha e domenit kohë është e tille që modeli all-pol është gjithmone i qendrueshëm .

$-$*ParashikimiLinearPonderues* (*WLP*)

Ashtu si parashikimi linear tradicional (LP) ,mostra $x[n]$ është llogaritur apo vlerësuar si kombinim linear ip mostrave të mëparshme :

$$\hat{x}[n] = -\sum_{i=1}^{p} a_i x[n-i] \qquad (2.6)$$

Ku koeficientet $a_i \epsilon \Re$ .Gabimi parashikues $e_n(a)$, mbetja , përcaktohet si :

$$e_n(a) = x[n] - \hat{x}[n] = x[n] + \sum_{i=1}^{p} a_i x[n-i] = a^T x[n] \quad (2.7)$$

Ku $\boldsymbol{a} = [a_0 a_1 \dots a_p]^T$ dhe $a_0 = 1$ dhe $x[n] = [x[n] \dots x[n-p]]^T$.Parashikimi i fuqisë së gabimit $E(a)$ në metodën WLP jepet nga barazimi

$$E(\boldsymbol{a}) = \sum_{n=1}^{N+p} (e_n(\boldsymbol{a}))^2 w_n = \boldsymbol{a^T} \left(\sum_{n=1}^{N+p} w_n x[n] x^T[n] \right) \boldsymbol{a} = \boldsymbol{a^T R a} (2.8)$$

Ku w_n është pasha apo koeficienti që i vendoset në mostrën e $n-$ te , N është gjatesia e sinjalit x[n] , dhe $\boldsymbol{R} = \sum_{n=1}^{N+p} w_n x[n] x^T[n] (2.9)$.Ky është një problem i kufizuar minimizimi

$$\min E(\boldsymbol{a}) mekushtqe \boldsymbol{a^T u} = 1$$

Ku $\boldsymbol{u}$ është një vector i percaktuar si $\boldsymbol{u} = [1\ 0\ 0 \dots 0]^T$. Shohim që matrica e autokorrelacionit është një matrice e peshuar ndryshe nga matrica e metodës së parashikimit linear tradicional.

Matrica **R** është simetrike por jo Toeplitz,për shkak të procesit të peshimit .Megjithate ajo është e përcaktuar pozitivisht dhe kjo e bën problemin e minimizimit konveks.Duke përdorur shumëzuesit e Lagranzhit ,tregohet që**a** kenaq ekuacionin linear

$$\boldsymbol{Ra} = \sigma^2 \boldsymbol{u}$$

ku $\sigma^2 = \boldsymbol{a^T R a}$është fuqia e gabimit .Së fundmi modeli WLP all-pol përcaktohet si $H(z) = {}^{1}/_{A(z)}$ ku $A(z)$ eshte Z-transformimi I vektorit **a** .

$-\boldsymbol{FunksioniPonderuar}$

Funksioni I ponderuar I domenit kohe w_n, eshte pika kyce e ted y metodave WLP dhe SWLP . Ne [41] ,funksioni ponderuar zgjidhet qe te jete Energji me Kohe te Shkurter (STE).

$$w_n = \sum_{i=0}^{M-1} x_{n-i-1}^2 = \sum_{i=n-M}^{n-1} x_i^2$$

Dmthenë pesha w_n e funksionit STE per mostren e n-te ,x_n eshte llogaritur duke marre energjine e sinjalit ten je dritareje te levizshme te *M* mostrave te meparshme.

Ku M është gjatësia e dritares STE. Përdorimi i dritares STE mund të justifikohet si më poshtë :

Funksioni STE thekson apo nënvizon mostrat e fjalës me amplitudë të madhe .Ajo njihet për faktin që aplikimi i analizës se parashikimit linear ne mostrat e fjales me interval faze te mbyllur glottal do te sjelle pergjithesisht nje paraqitje spektrale te fuqishme te traktit vocal ,e thene ndryshe theksojme qe funksioni peshe STE , perdoret nga modeli WLP per te theksuar zona shume energjike te kornizes se fjales .Ne keto zona ,energjia e zhurmes ne lidhje me ate te sinjalit ,eshte me e ulet, vecanerisht ne rastin e zhurmes stacionare shtese.Pervec kesaj per nje fjalim te shprehur shtese qe formohet ,kufijte e funksionit STE perkojne me fazen e mbyllur te ciklit glottal ,kur formants jane gjithashtu te lehta per tu shquar apo dalluar.

Nga matrica e autokorrelacionit te WLP ne ekuacionin (2.9),mund te shihet se nqse ponderimet i mostres eshte percaktuar per nje vlere konstrante,modeli all-pol I krijuar saktesisht perputhet me modelin konvencional te parashikimit linear.Perkatesisht duke rritur parametrin M ,shkaktohen perafrime te sjelljes se modeleve WLP ose SWLP , me majat me te theksuara spektrale .Ne te kundert duke perdorur nje dritare te shkurter STE,modeli I krijuar , rezulton ne vleresim spectral me te zbutur se sa modeli qe thekson pjeset e shkurtra me te fuqishme te kornizes se fjales.

$-\boldsymbol{Stabiliteti}$

Megjithate metoda WLP me dritare STE nuk siguron stabilitetin e modelit all-pol.Prandaj ne [42] ,nje formule per nje funksion ponderimi pergjithesues qe do te perdoret ne WLP eshte zhvilluar per te siguruar stabilitetin.Matrica **R** e autokorrelacionit e shprehur si me siper mund te shprehet ne trajten

$$\boldsymbol{R} = \boldsymbol{Y^T Y} (2.11)$$

Ku $\boldsymbol{Y} = [\boldsymbol{y_0 y_1} \dots \boldsymbol{y_p}] \boldsymbol{\epsilon} \boldsymbol{\Re}^{(N+p)x(p+1)}$ dhe $y_0 = [\sqrt{w_1}x[1] \dots \sqrt{w_N}x[N]\ 0 \dots 0]^T$

Vektoret shtylla jepen nga : $\boldsymbol{y_{k+1}} = \boldsymbol{B y_k} k = 0,1, \dots, p-1$ (2.12)

Ku

$$\boldsymbol{B} = \begin{bmatrix} 0 & 0 \ldots . & 0 & 0 \\ \sqrt{w_2/w_1} & 0 \ldots & 0 & 0 \\ 0 & \sqrt{w_3/w_2} \ldots & 0 & 0 \\ & \vdots & & \\ 0 & \ldots \ldots \sqrt{w_{N+p}/w_{N+p-1}} & & 0 \end{bmatrix}$$

Kështu ,para formimit te matrices **Y** , elementet e diagonales se dyte te matrices **B** ,përcaktohen per cdo i=1,2,…,N+p-1 në trajten

$$\boldsymbol{B}_{i+1,i} = \begin{cases} \sqrt{w_{i+1}/w_i} if w_i \leq w_{i+1} \\ 1 \qquad\qquad if w_i > w_{i+1} \end{cases}$$

Së fundmi themi që metoda WLP ,duke perdorur matricen **B** do të quhet modeli i$ParashikimitLinearPonderuesStabiloseIQendrueshem$dhe stabiliteti I modelit all-pol eshte siguruar .

Në disa aplikime ,si ne kodimin audio apo sintetizimin e fjales ,stabiliteti i modelit është thelbësor .Megjithatë ,në nxjerrjen e tipareve apo karakteristikave të njohjes së fjalës ,vetëm nje vleresues i spektrit i paraqitur nga modeli është nevojshem .Prandaj ne duhet të nxjerrim përgjigjen në frekuence të filtrit sintetizues LP , H(z), të dhënë si me sipër ,për marrjen e koeficientëve të WLP .

Filtri analizues LP është inverse i filtrit sintetizues ,

$$A(z) = \frac{1}{H(z)} = 1 - \sum_{k=1}^{p} a_k z^{-k}$$

Filtri analizues është një filter FIR dhe për këtë arsye ai është i qëndrueshem, dhe magnitude e spektrit te tij mund te llogaritet me ane te FFT te koeficienteve te filtrit.Përgjigja e magnitudës së filtrit sintetizues eshte inverse i ketij spektri .Në qoftë se sistemi nuk siguron qëndrueshmërine ,magnitude e spektrit te filtrit analizues ,mund te jete zero ,e cila do te coje ne vlera te pafundme ne vleresimin perfundimtar te zhvillimit spectral.Per te shmangur kete te gjithe vlerat e magnitudes se spektrit analizues qe e tejkalojne sasine e dhene nen maksimalen mund te lidhen me ate vlere.

Vlera e funksionit të transferimit jep amplitudë si tipar apo karakteristike e sinjalit të fjalës.Informacioni që marrim nga metoda LP nuk është i detajuar krahasuar me metodat e tjera.Ploti që i përket LP është i bute (smooth),në këtë mënyre ajo nuk zbulon shumë informacion.Kjo ndodh sepse ai nuk jep një tabllo të qartë të ndryshimit te funksionit apo karakteristikës ndërmjet dy vlerave të ndryshme të frekuencave. WLP na jep një informacion më të mirë në lidhje me karakteristikat.Nga plotet mund të shohim që karakteristikat janë me te hollësishme ,sepse ploti i WLP është më kodrinor dhe kuptohen më mirë se si ndryshojnë karakteristikat në frekuenca të ndryshme.

Kapitulli III

3.1 Gjuha Shqipe dhe Sinteza e tekstit

Gjuha shqipe, si një degë e veçantë në trungun e gjuhëve indoeuropiane përmban një grumbull rregullash të shkruarjes dhe leximit. Një pjesë e këtyre karakteristikave kanë ngjashmëri me disa gjuhë të tjera, sidomos atyre sllave dhe në veçanti në aspektin e leximit . Derisa nga aspekti fonetik dhe gramatikor ka një afërsi më të madhe me disa gjuhë tjera, si ato romane (italiane, spanjolle, portugeze) si dhe gjuhën greke .

Mënyra unike e leximit të simboleve (grafemave), rrjedhimisht edhe e fjalëve (me pak përjashtime) është një karakteristikë me shumë përparësi, sidomos për aspektin teknik të digjitalizimit. Në këtë drejtim, gjuha shqipe është më afër me gjuhën greke, gjuhët sllave si dhe me gjuhët fino-hungareze .

Në shumë gjuhë, problem në vete paraqet mënyra e leximit (si p.sh. më gjuhën angleze), dhe nga këndvështrimi i gjenerimit të të folurit paraqet vështërsi të mëdha. Kur kësaj i shtohen edhe elementet tjera, si leximi i vlerave numerike, datave apo edhe akronimeve, kjo karakteristikë i bënë gjërat edhe më të vështira .

Një karakteristikë kompleksiteti që i shtohet gjuhës shqipe janë simbolet diakritike që i përdorë alfabeti ynë: shkronjat ë, ç, sidomos duke u bazuar në faktin që shumë njerëz gjatë shrimit i tejkalojnë këto simbole duke i zëvendësuar me e dhe c, respektivisht. Kjo bënë që gjatë konvertimit shumë fjalë të lexohen ndryshe nga ajo që e ka menduar shkruesi

Shumë nga të përbashkëtat e gjuhës shqipe me gjuhët tjera mund të shfrytëzohen në modelin ASR për sintezën e të folurit.

Derisa mënyra e leximit të teksteve në shqip është dukshëm më e thjeshtë, krahasuar me shumë gjuhë tjera, gramatika dhe intonacioni është më i komplikuar . Mënyra unike e të lexuarit të çdo simboli (grafeme), pa marrë parasysh pozitën ku ndodhet dhe me çfarë shoqërohet (me përjashtim të bigrameve – grafemave dysimbolëshe) është një përparësi shumë e madhe edhe gjatë përpilimit të algoritmeve për krijimin e të folurit artificial.

Gjuha shqipe megjithese flitet nga shume njerez deri me tani shume pak burime akustike (baze e te dhenave te te folurit)dhe burime gjuhesore (korpuset tekst) jane perftuar per te zhvilluar nje system te pafrenuar te vazhdueshem te njohjes se fjales. Paradigma e trajnimit Baum-Welch kerkon klipe fjalimi audio se bashku me transkriptimet tekstuale te tyre ne menyre qe te vleresoje parametrat e modelit .Keshtu bazat e te dhenave te te folurit jane burime kritike se bashku me karakteristikat e tyre te tille si numri I oreve te fjales ,numri I folesve etj ,ne zhvillimin e nje sistemi te njohjes se fjales. Sic permendem me lart gjuha shqipe nuk ka ne dispozicion burime te te folurit as komercial as free .Per me teper pavaresia e gjuhes nuk mjafton per te nenkuptuar burimet nga nje numer I madh I folesve.

Arkitektura e shumicës së sistemeve ARS moderne komerciale është e bazuar në sintezën përmes bashkimit të segmenteve zanore (akustike) të regjistruara (incizuara) paraprakisht dhe të ruajtura si fajla akustik. Kjo është e njohur si Concatenative Synthesis .

Në përgjithësi, sinteza vargëzuese rezulton me gjenerim natyral të të folurit. Megjithatë, ekzistojnë edhe dallime në mes të të folurit natyral dhe gjenerimit të të folurit përmes teknikave të automatizimit dhe segmentimit të formave të ndryshme valore. Përdoren tri nëntipe të sintezës së realizuar përmes kësaj teknike, që kanë të bëjnë me gjatësinë e segmenteve akustike.

Sinteza vargëzuese e bazuar në njësi (*Unit Selection Synthesis*) bazohet në selektimin e njësive nga një bazë e madhe të dhënash. Gjatë krijimit të bazës së të dhënave, nga të folurit e incizuar, ndahen njësitë përkatëse ose paralelisht në të gjitha njësitë si që janë: fonemat, dyfonemat, gjysëmfonemat, rrokjet, apo fjalët. Ndarja bëhet duke shfrytëzuar programe të veçanta për njohjen e të folurit (*speech recognizer*) ose me intervenime manuale dhe zakonisht të shoqëruara edhe me paraqitje vizuale, siç janë: format valore (*waveform*) dhe spektrogramet (*spectogram*).

Sinteza e bazuar në dyshe fonemash (*diphone*) dallohet me faktin që përmban bazë të vogël me fajlla akustik. Numri i njësive të këtilla, të njohura si “diphone” varet nga fonetika e gjuhëve të veçanta dhe është relativisht i ndryshëm, (p.sh. gjuha gjermane ka rreth 2500 difone, ajo spanjolle rreth 800, etj.). Gjuha shqipe përdorë diç më pak se 1200 difone (gjithsej ka 1296 difone, prej të cilave disa nuk përdoren, ngase nuk kanë kuptim). Kështu, përmes këtij numri të njësive gjenerohet të folurit, duke shfrytëzuar teknika të procesimit digjital të sinjaleve siç është kodimi linear prediktiv .

Kualiteti i sintezës së këtillë është më i ulët, por është më natyral në krahasim me sintezën e tipit formant. Manipulimi me databazën është më i lehtë.

Sinteza e të folurit për destinime specifike bën bashkimin e fjalëve ose shprehjeve pët të gjeneruar të folurit. Kjo teknikë zakonisht shfrytëzohet në aplikacione specifike, ku numri i fjalëve dhe i shprehjeve është i kufizuar, si p.sh. te kalkulatorët, ora, koha, orari i trenave, etj. Teknologjia e implementuar është e thjeshtë dhe sisteme të këtilla ka me vite që janë në shfrytëzim komercial, p.sh. ora për fëmijë që flet.

Sinteza e këtillë ofron kualitet të lartë të gjenerimit natyror të të folurit, ngase realizon kombinimin e njësive origjinale të incizuara në bazën akustike të të dhënave, të cilat në fakt janë fjalë ose shprehje. Andaj, sinteza e këtillë nuk mund të shfrytëzohet për destinim të përgjithshëm, por vetëm në domene specifike.

3.2 Kërkesat që i parashtrohen një sistemi funksional të njohjes së fjalës

Një sistem për konvertim të te folurit në tekst është një sistem kompleks që duhet të përmbajë pjesët për futjen dhe analizën gjuhësore, si dhe pjesët për përpunimin digjital për sintezën përfundimtare të të folurit . Të folurit e njeriut përbëhet nga fjalët, të cilat përbëhen nga segmentet elementare zanore që njihen si *fonema*. Numri i fonemave bazë është i ndryshëm për gjuhë të ndryshme. Ndërsa, teksti i shkruar ndërtohet nga simbolet që ndryshe njihen si grafema. Për shkak se të folurit shpreh edhe shumë emocion dhe subjektivitet të folësit, shndërrimi i fonemes në grafeme nuk është unik dhe dallon nga fjala në fjalë, aq më tepër nga fjalia në fjali. Prandaj, ajo që në shikim të parë do të dukej e thjeshtë (shndërrimi direkt i fonemes në grafemen përkatëse), në të vërtetë duhet të ketë parasysh shumë parametra, në mënyrë që të folurit e gjeneruar të jetë i kuptueshëm dhe i natyrshëm .

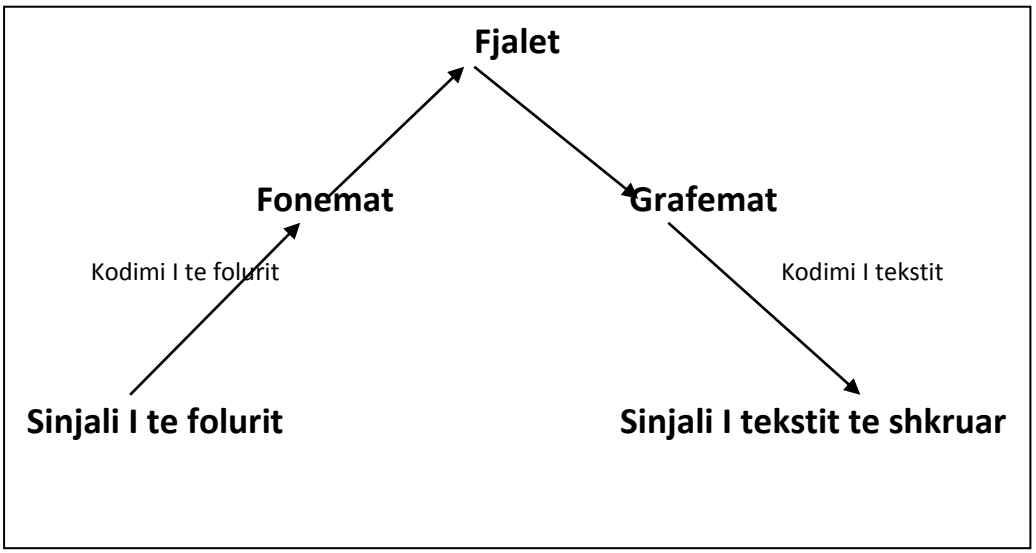


Figura 3.1: Modeli baze per konvertimin e te folurit ne tekst

Pjesa e gjuhës që merret me tingujt dhe përshkrimin e tyre, që njihet si fonetikë, ka rëndësinë më të madhe gjatë dizajnimit dhe projektimit të softuerëve për sintezën e të folurit.

Gjuha shqipe përmban 36 tinguj bazë që përshkruhen me 36 shenja të veçanta, të cilat njihen me emrin fonema. Simbolet grafike të fonemave janë ato që i njohim si shkronja. Nga këto fonema, 7 prej tyre përshkruajnë tinguj të zanoreve (a, e, ë, i, o, u, y) kurse 29 i përshkruajnë bashkëtinglloret (b, c, ç, d, dh, f, g, gj, h, j, k, l, ll, m, n, nj, p, q, r, rr, s, sh, t, th, v, x, xh, z, zh). Nga këto 9 janë shkronja dy-simboleshe (dh, gj, ll, nj, rr, sh, th, xh dhe zh), të cilat janë të njohura si "bigrames".Brenda grupit të shkronjave një simbol, 25 prej tyre janë alfabetit Latin , ndërsa 2 janë të përjashtuara dhe janëtë njohur si karaktere diakritike - shkronjat "ç" dhe "ë" . Shkronjat dy-simbol (bigrames) përbëjnë një sfidë në vetvete, gjatë konvertimit të tekstit në të folurit, sepse ata duhet të trajtohen si letër e vetme.
Në bazë të matjeve, frekuencat e përdorimit të shkronjavete gjuhës shqipe janë gjetur , siç jepen në figurën e mëposhtme .

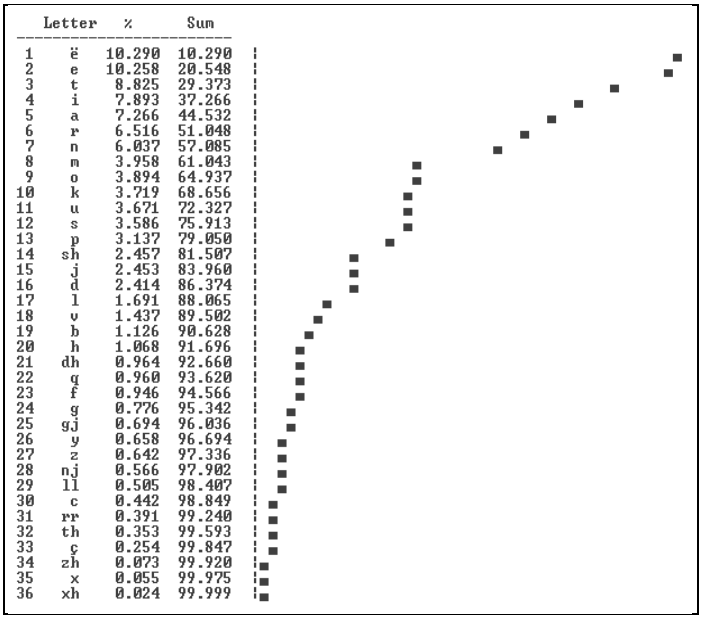


Figura 3.2 : Frekuencat e përdorimit të shkronjave të alfabetit shqip

Diagrama tregon se frekuenca më e madhe e përdorimit të kenë zanoret "e" dhe "ë". Por, në vendin e tretë duket bashkëtingëllorja "t", e cila është pjesë e fjalës "te", e cila është llogaritur si fjala më e shpeshtë në gjuhën shqipe. Pas kësaj, ashtu si në shumë gjuhë, si shkronja me frekuencë të lartë janë të vendosura dy zanoret e tjera: i dhe o.

Për të krahasuar shpeshtësinë e përdorimit të grafema të gjuhës shqipe me ato në gjuhë të tjera, në figurën më poshtë është treguar shpërndarja e frekuencave të përdorimit të grafema, në gjuhën shqipe.

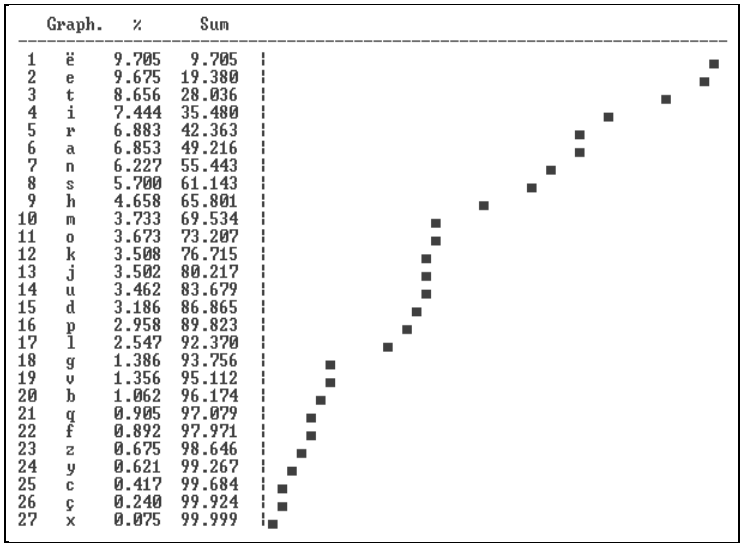


Figura 3.3 : Shpërndarja e grafemave të gjuhës shqipe

Një veçanti e gjuhës shqipe në krahasim me gjuhët e mëdha botërore është lexueshmëria e grafemave ashti siç shkruhen .Kjo veçanti ka të bëjë me një lexim të veçantë të grafemave pavarësisht se ku ato janë të vendosura.Pra çdo shkronjë lexohet në mënyrë unike .Përjashtim bëjnë 9 bigramat (shkronjat dysimboleshe) që në të shkruar përbëhen nga dy grafema dhe lexohen si një shkronjë e vetme.Për të eleminuar problemin e shkronjave me dy simbole ,para se të gjenerohen dosjet e zërit ,ato mund të zëvëndësohen me shkronja një simbolëshe ,që bën të mundur për ti trajtuar si shkronja të zakonshme .Një "problem" i veçantë në gjuhën shqipe janë dy diakritiket në gjuhën shqipe që nuk mund të gjenerohen direkt nga tastjera e kompjuterit .Këto janë grafemat "ë" dhe "ç".Grafema "ç" është shumë e afërt me "q",ndërkohë grafema "ë" nuk është e ngjashme me ndonjë grapheme tjetër në lexim .Për më tepër kjo është një shkronjë me frekuencë shumë të lartë në përdorim ,madje më të lartë se dhe shkronja "e".

Me kombinimin e fonemave formohen rrokjet, nga të cilat formohen fjalët, që paraqesin strukturën bazë të gjuhës. Rrokja është një grup fonemash të përqendruara rreth një zanoreje. Rrokjet janë të hapura kur mbarojnë me zanore (p.sh. *u-ra*) dhe të mbyllura kur mbarojnë me bashkëtingëllore (p.sh. *lis*). Një grup bashkëtingëlloresh pa zanore nuk mund të formojë rrokje.

Njësi tjetër që përdoret gjatë analizës së tekstit është edhe morfema. Morfema është njësi imagjinare akustike, që mund të jetë një shkronjë, dy shkronja, një rrokje apo një fjalë .

Ngritjet dhe uljet e zërit gjatë të folurit njihen si intonacione. Intonacioni mund të pësojë modifikime edhe në varësi nga shprehja e emocioneve dhe e ndjenjave të gëzimit, zemërimit, admirimit, shpresës, dyshimit, etj.

Shqiptimi i një rrokjeje të fjalës me forcë më të madhe se të tjerat quhet theks (*stress*). Rrokja që shqiptohet me forcë më të madhe quhet rrokje e theksuar, ndërsa të tjerat quhen të patheksuara. Në shqip, theksi mund të bjerë në njërën ose në tjetrën rrokje të fjalës. Përveç fjalës, edhe fjalia ka theksin e vet .

Intonacioni, theksi, mënyrat, rasat dhe karakteristikat tjera të fjalëve dhe fjalive janë me një rëndësi të madhe gjatë sintezës së të folurit, sidomos për natyralitetin e të folurit.

Studimi i strukturës së gjuhës mund të fillojë nga analiza e paraqitjes së fonemave, ngaqë ato paraqesin komponenten elementare të saj. Nga punimet që kanë të bëjnë me këtë sferë shihet që nuk mund të ketë një përgjigje përfundimtare lidhur me këtë çështje .

Megjithatë, gjetja e frekuencës së paraqitjes së fonemave në tërësinë e tekstit është një madhësi e vlefshme gjatë përpunimit , jo vetëm lidhur me gjenerimin artificial të të folurit por

edhe në çështje tjera, si telekomunikime, kriptografi, kodime të ndryshme, e të tjera . Sipas disa analizave të publikuara në punime nga analiza statistikore për gjuhën shqipe, janë nxjerrur disa përfundime me rëndësi, si p.sh. frekuenca e përdorimit të fonemave, grafemave, zanoreve, bashkëtinglloreve, digrafeve, etj.

Procesi i njohjes së gjuhë së folur është një prej proceseve më të vështira për t‟u kalkuluar nga ana informatike. Ndryshe nga aparati i dëgjimit të njeriut i cili i njeh mjaft mirë tingujt nëse ata janë shqiptuar qartë dhe nuk ka zhurma në mjedisin përreth, procesi i informatizuar i njohjes së fjalëve të folura vazhdon të mbetet një problem i pazgjidhur plotësisht. Edhe gjigandët informatikë industrial apo shkencorë vijojnë të provojnë teknika të reja në përmirësim të njohjes së gjuhë së folur. E megjithatë, edhe pse burimet dhe fondet nuk u mungojnë këtyre projekteve, rezultatet më të mira janë ende kryesisht të lidhura me identifikimin e fjalëve specifike. Për identifikimin e tog-fjalëshave apo fjalive rezultatet ende janë tepër të dobta. Shpesh herë mendohet se ende nuk është gjetur metoda optimale për të kryer saktë identifikimin e sinjalit zanor. Megjithatë, modeli akustik i zakonshëm i përdorur në njohësit e zërit është Modeli i Fshehtë i Markovit. “Modele të tjera janë të mundura, si ato të bazuara në rrjete artificiale neural apo në ndarjen dinamike të kohës.”
Nga ana tjetër duhet pasur parasysh se modeli akustik është vetëm hallka informatike e trajtimit të zgjidhjes së problemit të njohjes së gjuhës së folur.
Në fakt procesori akustik rrethohet nga njëra anë nga prodhuesi i sinjalit zanor dhe nga ana tjetër nga dekoduesi linguistik. Detyra konkrete e procesorit akustik është që duke përdorur përshkrimin e modelit akustik të krahasojë një përfaqësi të sinjalit hyrës me një përfaqësi të gjendjeve të regjistruara më parë për të gjetur qasjen e saktë. Në mënyrë që të merret një përfaqësi e saktë e sinjalit hyrës, metoda më e përdorshme do të qe ajo e paraqitjes së cepstrum-it të sinjalit. Pikërisht edhe gjetja e këtij cepstrumi quhet MFCC. Kjo metodë e përdorur fillimisht në aplikime sizmike duket sikur është metod më e saktë për përpunimin e sinjalit zanor. Pikërisht rezultati i kësaj metode do tëanalizohet në vijim (rezultati do të qe në rastin e përgjithshëm, një vektor me 39elementë për secilën nga zgjedhjet e njëpasnjëshme që i bëhen sinjalit – kampionimi Isinjalit – dhe do të qe paraqitja vektoriale e një subfoni).
Fillimisht u studiuan për këtë punë mënyrat e ndryshme sesi sinjali dixhital zanor mund të standardizohej në një mjet të procesueshëm më pas për Njohjen e Ligjëratës. E gjithë kjo ndërmarrje rezultoi në pranimin e modelit me koeficientë MFCC. Ky rezultat u arrit duke bërë krahasimin e literaturave të ndryshme të kohëve të fundit. Një analizë e detajuar e koeficientëve të ndryshëm të ekstraktuar nga sinjali zanor, paraqiti një dominim të

koeficientëve MFCC si rezultat final i përpunimit të sinjalit zanor dixhital, i cili paraqet bandat e ndryshme të frekuencave dominante në sinjal e që mund të përdoret si input për procesin e Njohjes së Ligjëratës.

Komuniteti ASR këtu vendoset duke përdorur dy teknika të nxjerrje së tipareve të ngjashme dhe dominuese,sic janë teknika MFCC dhe ajo LPC (parashikimi linear perceptual).Të dyja teknikat përdorin degjimin si deformim apo shtrembërim të spektrit të shkurtër kohor të sinjaleve akustike , si dhe japin rezolucion spectral më të lartë në frekuenca të ulta .Këto karakteristika themelore kanë për qëllim të imitojne pjesët statike te spektrave ashtu sic është përceptuar nga njerëzit.Pervec nga karakteristikat statike ,ndryshimet në kohë të paraqitura nga parametric delta dhe i përshpejtimit luajnë një rol të rëndesishëm në modelimin e evolucionit të fjalës.Si pasojë shumica e sistemeve të njohjes së fjalës sot përfshijne tiparin apo funksionin delta si një add-on per parametrat statike.Hapat që mbulojnë nxjerrjen e këtyre tipareve standarte janë parapërpunimi,analiza e bankës së filtrave dhe gjenerimi i cepstrumit.

3.3 Burimet e nevojshme për të ndërtuar një sistem të njohjes

3.3.1 Formalizimi i njohjes

Procesi i njohjes automatike të fjalës është përkthimi i fjalëve të folura në tekst.Kjo detyrë fjalim-ne –tekst mund të karakterizohet apo mund të shihet në një kuadër probabilistik.Teoria e probabilitetit dhe statika sigurojnë gjuhen matematikore për të analizuar dhe përshkruar sistemet ASR .Detyra fjalim-ne-tekst mund formulohet në një menyre probabilitare :

Cili është rendi(sekuenca) me i mundshëm i fjalëve W^* në një gjuhë të caktuar L duke patur parasysh shprehjen fjalim X ?

Paraqitja formale që përdor funksioni arg max , në mënyrë që të zgjedhim argumentin i cili maksimizon probabilitetin e fjalës sekuence është :

$$W^* = arg \max_W p(W/X) \qquad (5.1)$$

Ekuacioni (5.1) thekson sekuencen më të mundshme të fjalëve si një sekuencë me probabilitet posterior më të lartë duke patur parasysh shprehjen fjalim.Ky probabilitet posterior është llogaritur duke përdorur rregullat e Bayesit ,kështu që sekuenca me e mundshme e fjalës bëhet :

$$W^* = arg\max_W \frac{p(X/W) * p(W)}{p(X)} \qquad (5.2)$$

p(X) probabiliteti i fjalës është i pavarur nga sekuenca e fjalëve W,dhe mund të injorohet.Problemi i njohjes reduktohet thjesht në :

$$W^* = arg\max_W p(X/W) * p(W) (5.3)$$

Ekuacioni (5.3) vë në dukje dy mandate që mund të vlerësohen direct a) probabiliteti apriori i sekuencës së fjalës p(W) dhe b) probabiliteti i të dhënave akustike duke patur parasysh sekuencën e fjalës p(X/W).Faktori i parë mund të vlerësohet duke përdorur një model gjuhësor ,ndërkohe që faktori i dyte mund të vlerësohet me ndihmën e një modeli akustik.Të dy modelet mund të ndërtohen në mënyrë të pavarur ,por do të përdoren se bashku në mënyrë që të dekodohen apo deshifrohen të dhënat e të folurit ,sic tregohet në ekuacionin 5.3.Arkitektura e përgjithshme e një ASR tregohet në figurën e mëposhtme .Në të mund të shohim që procesi i njohjes së të folurit pershkruhet në dy faza themelore : a) nxjerrja e informacionit të dobishem nga sinjali i fjalës dhe b) ngjeshja e këtyre paraqitjeve apo përfaqësimeve për trasmetimin efikas dhe magazinimin.

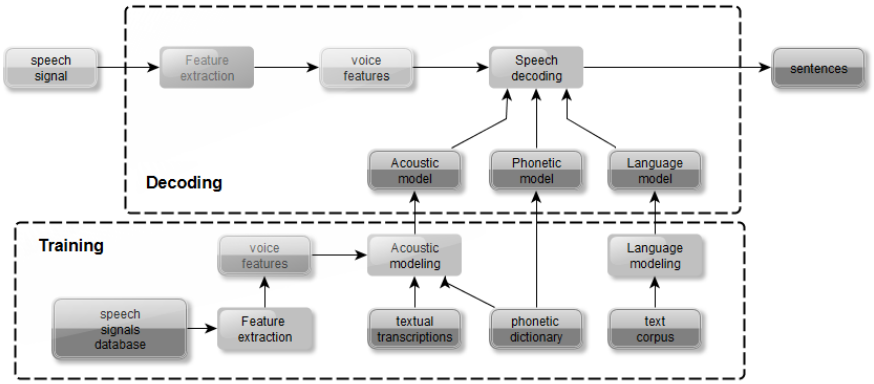


Figura 3.4 : Burimet e nevojshme per ndertimin e ASR

Përveç modelit gjuhësor dhe atij akustik të permendura më lart ,arkitektura e përgjihshme e ASR përfshin gjithashtu një model fonetik.Qëllimi i tij është të lidh modelin gjuhësor me atë akustik.Figura e mësiperme tregon gjithashtu se në një fazë të hershme sistemi kryen një nxjerrje tiparesh .Ky bllok ka rolin e nxjerrjes së karakteristikave të veçanta akustike te cilat përdoren më tej për të krijuar modelin akustik .Si pasoje i njëjti bllok i nxjerrjes së tipareve është përdorur në procesin e deshifrimit .Më poshtë do të përshkruajme dhe analizojme disa blloqe të tjera nga arkitektuara e ASR.

3.4 Modeli Gjuhësor

Modeli gjuhësor (gramatika) është përdorur në fazën e deshifrimit për të përshkruar sa është e mundshme në një kuptim probabilitar ,që një sekuence e simboleve të gjuhës mund të mund të shfaqet në një sinjal të fjalës.Një model gjuhësor statistikor cakton nje probabilitet te nje sekuence prej n fjalesh $P(w_1, w_2, \dots, w_n)$ me anë të shperndarjes probabilitare .Qëllimi kryesor i gramatikës është që të vlerësojë probabilitetin që një sekuence fjaleve $W = w_1, w_2, \dots, w_n$ është një fjali e vlefshme ne gjuhen e hulumtuar .Probabiliteti i sekuences se ketyre fjaleve ndihmon modelin akustik ne procesin e vendimarrjes.

Me fjalë të tjera ,një model gjuhesor është përdorur për të kufizuar kërkimin e fjalëve.Ai përcakton cilat fjalë mund te ndjekin fjalet e njohura me pare dhe ndihmon per te kufizuar ne menyre te konsiderueshme procesin duke hequr fjalet qe nuk jane te mundshme.Probabiliteti i sekuences se fjaleve $W = w_1, w_2, \dots, w_n$ mund te zberthehet si vijon:

$$p(W) = p(w_1, w_2, \dots, w_n) = p(w_1)p(w_2|w_1)p(w_n|w_1, w_2, \dots, w_{n-1})$$
$$= \prod_{i=1}^{n} p(w_i|w_1, w_2, \dots, w_{i-1}) \qquad (5.4)$$

Ku $p(w_i|w_1, w_2, \dots, w_{i-1})$ është probabiliteti qe w_i të ndodhë ,duke patur parasysh që sekuenca e fjaleve $w_1, w_2, \dots, w_{i-1}$ ka qene e pranishme me pare.

Tani detyra për të vlerësuar probabilitetin e një sekuence fjalësh është e ndarë në disa detyra të vlerësimit të probabilitetit të një fjale për të dhënë nje histori te fjalëve të mëparshme .Në ekuacionin 5.4 një zgjedhje e tillë varet nga e gjithë pjesa e histories së inputit.Për një fjalor të madhësive të ndryshme ka histori të ndryshme dhe në mënyrë që të përcaktohet plotësisht duhen vlerësuar vlerat.Në realitet ,probabilitetet nuk mund të bëjne vlerësimin për çdo vlerë të moderuar ,pasi shumica e historive eshte unike ose kane ndodhur vetem disa here.Një zgjidhje praktike është që të supozojmë që varet vetëm nga disa klasa ekuivalence.Klasa e ekuivalences mund të bazohet thjesht në disa fjalë të mëparshme .Kjo çon në një model ghuhësor n-gram.Trigram është një rast i vecante dhe është provuar që është shumë i fuqishëm ,pasi shumica e fjalëve kanë një varësi të fortë nga dy fjalë të mëparshme.

3.4.1 Modelet N-Gramsh

Modeli n-gram ,i cili karakterizon marrëdhenien e fjalës Brenda një hapesire prej n-fjalësh është një paraqitje shumë e fuqishme statistikore e një gramatike.Efektivitetin e saj në ndërtimin e një kërkimi fjale është vërtetuar fuqimisht nga loja e famshme eClaude Shannon e cila konsistonte në konkursin midis një njeriu dhe nje kompjuteri.Në këtë konkurs te dy si njeriu dhe kompjuteri u pyeten në mënyrë të njëpasnjeshme të mendonin një fjalë tjetër në një fjali të rastësishme.Njeriu mendoi në përvojën e gjuhës amtare ,ndërkohe që kompjuteri i bazoi pergjigjet e tij ne parimin e përgjasise maksimale,duke përdorur statistikat e fjalëve të akumuluara.Ky eksperiment tregoi se kur ,n, numri i fjalëve të mësipërme kalon 3 ,kompjuteri kishte më shumë gjasa që të jepte një mendim me të mirë të fjalës tjetër ne një fjali se sa njeriu.Aktualsht modelet n-gram janë të domosdoshme në sistemet e njohjes së fjalës në rastin e fjalorëve te medhenj.

3.5 Modeli fonetik

Në kontekstin e sistemit të njohjes së fjalës së vazhdueshme ,modelet akustike nuk modelojnë fjalët e gjuhës burim në mënyrë të drejtpërdrejtë ,por në mënyre te terthorte .Për sistemet e njohjes së fjalës së vazhdueshme të fjalorëve të medhenj,ku fjalor i madh ne pergjithesi do të thotë se sistemet kanë përafersisht 5000deri ne 60000 fjalë,është e vështire të ndërtohen modele gjithe-fjalësh .Termi i vazhdueshem i referohet faktit se këto fjalë janë të drejtuar njekohesisht natyrshëm dhe jo të izoluara ,ku cdo fjale do te paraprihet dhe te ndiqet nga nje pause.Qëllimi i modelit fonetik është që të lidh modelin akustik ecila vlereson probabilitetet akustike të fonemave ,me modelin gjuhesor ,e cila vlereson probabilitetet e sekuencave te fjaleve.Komponenti analiza fonetike konverton tekstin e perpunuar ne sekuencen perkatese fonetike.Me fjale te tjera ,fjalori fonetik eshte nje instrument gjuhesor ,e cila ben korrespondencen midis forms se shkruar dhe forms fonetike te fjaleve ne nje gjuhe burim.Kjo ndiqet nga nje analize prosodike per te bashkengjitur apo bashkuar regjistrin (pitch) perkates dhe kohezgjatjen e informacionit ne sekuencen fonetike.Ne gjuhesi ,prosodike eshte stresi,ritmi dhe intonacioni i te folurit.Prosodikja ose metrika mund te tregoje disa karakteristika te folesit ose shprehje,si gjendjen emocionale te folesit ose formen e fjales (deklarate,komande ose pyetje)oseprania e ironies apo sarkazmes dhe shume elemente te tjere te gjuhes qe nuk und te kodohen nga gramatika ose nga zgjedhja e fjalorit.

Edhe pse nje fjalor i ndërtuar me dorë do të siguronte një fonetike të përsosur ,kjo detyrë do të kërkonte shume kohe kur do të ndërtonim sistemin e njohjes së fjalës së vazhdueshme .Për më tepër kjo do të kërkonte një zotërim të mirë të gjuhës respektivisht.

3.6 Modeli Akustik

Modelet akustike i referohen kryesisht paraqitjes së njohurive rreth fonetikës ,akustikes,prononcimeve të ndryshme ,gjinisë dhe dallimeve dialektore mes folesve,ndryshueshmërisë së mjedisit dhe kështu me rradhë.Një sistem i njohjes së fjalës i cili mund të aplikohet në një numër të madh folësish ,papasur nevojë për ti trajtuar individualisht cdo njerin prej tyre quhet sistem i pavarur nga folesi.Një sistem i tillë është i bazuar ne disa algoritme ,me qëllimin final të krijimit të fjalës dhe modeleve të zërit referues,te cilat mund të përdoren në një gamë të gjërë të folësve dhe thekseve.Në fazat më të hershme ,këto modele jane karakterizuar nga nje qasje intuitive ,por gradualisht evoloi ne modele rigoroze statistikore.

Popullariteti dhe përdorimi i modeleve të fshehura të Markovit si bazë kryesore për njohjen automatike të të folurit ka mbetur e pandryshueshme gjate dy dekadave te fundit.HMM është sot metoda më e preferuar e njohjen e të folurit kryesisht për shkak të rrjedhes së qëndrueshme të përmirësimit të teknologjise .Një tjetër arsye pse HMM janë të njohura është se ato janë të trajnuara në mënyrë automatike dhe të lehta ne perdorim .Modelet HMM mund te ofrojne nje menyre efikase per te ndertuar modele te denja parametrike dhe gjithashtu te perfshijne parimin e programimit dinamik ne thelbin e saj per unifikimin e segmentimit te modelit dhe klasifikimin e modelit te sekuencave te te dhenave ne varesi nga koha .Supozimi themelor i HMM eshte se mostrat e të dhënave mund të karakterizohen si një process rasti parametrik ,dhe parametrat e procesit stokastik mund të vlerësohen në një kuader të saktë dhe të përcaktuar mirë.HMM është bërë një nga metodat më të fuqishme statistikore per modelimin e sinjalit të fjalës.Parimet e saj janë përdorur me sukses në njohjen automatike të folurit,në gjurmimin e formanteve dhe regjistrit të zërit,të kuptuarit e gjuhës së folur .

3.6.1. Modeli i Fshehtë i Markovit dhe përdorimi i tij në modelet akustike

Forma më e thjeshtë e modeleve të Markovit janë modelet e zinxhirëve tëMarkovit, të cilat mund të përdoren për përshkrimin statistikor të simboleve dhesekuencave të gjendjeve. Në të

ashtuquajturat modele të fshehta të Markovit(HMM) koncepti i një sekuence gjendjesh, i cili modelohet statistikisht, zgjerohet medaljet specifike të gjendjes së modelit."[8] Nga ky përkufizim i modelit të Markovitnënkuptohet fakti se ky lidhet ngushtë me një sekuencë gjendjesh rrjedhojë të ngjarjeve tërastit (mbi të cilat kemi ngritur të dhëna statistikore). Nga ana tjetër duhet pasur parasyshse informacioni qe vjen nga procesuesi i sinjalit zanor është një varg i panjohur që mëparë duke pasur parasysh se ai varet në mënyrë të drejtpërdrejtë prej folësit. Sigurisht qëfolësi kryen jo një proces të thjeshtë rasti por shkronjë pas shkronje ai artikulon një fjalëqë gjendet në fjalor ose një variacion mbi fjalën e fjalorit; si dhe fjalë pas fjale krijon njërrjedhë logjike të lidhur me semantikën e shprehjes. Fjala vjen nën rrjedhën e disanënzërave (subfoneve). Secili nga këto kampionime nënzërash duhet të tregojë se cilënprej fonemave bazë të gjuhës, e më pas cilën shkronjë – siç ndodh në rastin e gjuhësshqipe që çdo shkronjë përfaqësohet nga një fonemë duhet përzgjedhur. Procesi është relativisht i vështirë dhe jo kaq i drejtpërdrejtë dhe gjetja e një modeli të saktë lidhet medisa algoritme që duhet të veprojnë në grup për të dhënë një rezultat real.

"Ajo që në fund duhet të rezultojë në dalje të këtij Modeli të Fshehtë Markovi dotë qe duke mbështetur rregullat e Bajesit, të kishim si rezultat se sekuenca më e mirë efjalëve është ajo që maksimizon produktin e dy faktorëve, një model gjuhësor tëmëparshëm me një ngjarje rasti akustike."[9]Shprehja matematikore e asaj që shprehëm më lart do të qe (ku **L** është gjithëfjalori me fjalët dhe O është bashkësia e vrojtimeve të rastit):

$\boldsymbol{W} = argmax\ P(O|W)\ P(W) W \epsilon \boldsymbol{L}$

Ky më saktësisht është modeli i kanalit me zhurmë. Le të fillojmë dalëngadalë tëshqyrtojmë elementët që marrin pjesë në këtë ekuacion.

$Q = q_1, q_2, \dots, q_n$ – janë N gjendje të ndryshme (Këto N gjendje të ndryshme jodomosdoshmërisht përfaqësojnë një shkronjë. Ato përfaqësojnë një subfon ose më saktë ethënë një pjesëz e sinjalit të përftuar nga një shkronjë. Kjo lidhet me mënyrën e shprehjessë çdo shkronje si dhe në mënyrë të drejtpërdrejtë me ndarjen e njëjtë që i bëhetzinxhirëve të sinjalit që vijnë për fjalët – në frame – ndarje kjo që nuk mund të shmangetpasi sistemi endë nuk e di a ka shkronja apo thjesht tinguj – klithma – në sinjal.)

$A = a_{11}, a_{12}, \dots, a_{1n}$

$a_{21}, a_{22}, \dots, a_{2n}$

…

$a_{n1}, a_{n2}, \dots, a_{nn}$ – janë të gjitha probabilitetet e kalimeve nga një gjendje i në njëgjendje j të tilla që shuma e daljeve nga gjendja i drejt të gjitha gjendjeve j do të qe 1 (praprobabiliteti të

qe 100%).O = o_1, o_2, … , o_T – T vrojtime të nxjerra nga një fjalor F. Këto do të qenë më saktëshkronjat e fjalorit fonetik ose fonemat.

B = $b_i(o_t)$ – janë ato që quhen mundësitë e vrojtimit apo probabilitetet e emetimit të cilatjanë probabilitetet e emetimit të një vrojtimi ot nëse jemi në gjendjen i. Tregon sa është mundësia që në një frame të caktuar (në frame-in e i-të) të kemi kapur shkronjën o_j.q_0, q_F – janë statuset fillestare dhe finale të cilat kanë edhe këto një varg brinjësh qëlidhin këto gjendje me secilën prej n gjendjeve (përkatësisht: a_{01}, a_{02}, ... ,a_{0n} dhe a_{F1}, a_{F2},... , a_{Fn}).

3.6.2 Struktura HMM

HMMs janë shumë të njohura në fushën e njohjes së fjalës ,për shkak të avantazheve që ato ofrojnë.Në krahasim me metodat e thjeshta Markov ,në rastin e HMMs nuk ka bijeksion midis gjendjes dhe outputit.Kjo ofron një fleksibilitet më të madh dhe përputhet në mënyre perfekte me sinjalin e fjalës,në te cilin e njejta foneme mund të ketë kohezgjatje të ndryshme ne varësi te rastit .Një model i fshehur i Markovit eshte nje process stokastik ,i cili modelon ndryshueshmerine e brendshme te sinjalit te fjales,dhe strukturen e gjuhes se folur ne nje kornize te qendrueshme te modelit statistikor.HMMs jane makina me gjendje te fundme probabilitare,te cilat mund te kombinohet per te marre modele sekuencash fjalesh nga njesi me te vogla.Ne detyren e njohjes se fjales me shkalle te larte fjalori,sekuencat e fjaleve jane ndertuar ne menyre hierarkike nga modelet e fjales te cilat nga ana tjeter jane ndertuar nga modelet nen-fjale me ndihmen e nje shqiptimi te fjalorit.Per rezultate te mira te njohjes ,keto modele nen-fjale duhet te jene modele te telefonit ne kontekstin e varesise.

Sipas natyres se sa ,nje sinjal i fjalës është ndryshueshem,për shkak te variacioneve te shqiptimit apo faktoreve te mjedisit.Kur e njejta fjale eshte thene nga disa foles,sinjalet akustike mund te jene te ndryshme edhe pse struktura themelore gjuhesore mund te jete e njejte.HMM perdor zinxhiret e Markovit per te krijuar strukturen gjuhesore dhe nje sere shperndarjesh probabilitare per te shenuar ndryshueshmerine ne realizimet akustike te tingujve te te folurit.Dhenia e nje koleksioni te mjaftueshem e variacionit te fjaleve me interes,mund te marrim bashkesine me te pershtatshme te parametrave qe percaktojne modelin korespondues ose modelet,perms nje metode efikase vleresimi sic eshte algoritmi Baum-Welch.Ky vleresim i parametrave mund te perkthehet perms trajnimit dhe mesimit te sistemit.Ne fund ,modeli resultant duhet te jete ne gjendje te tregoje nese nje shprehje e panjohur eshte me te vertet nje realiz I fjales qe paraqitet ne model.Modeli i fshehur i

Markovit paraqet nje process jo-te percaktuar qe gjeneron simbolet e vezhgimit te outputit ne ndonje gjendje te dhene.Pra vezhgimi eshte nje funksion probabilitar i gjendjes.Ai mund te shihet sin je process stokastik dyfish i ngulitur me nje process themelor stokastik (sekuenca gjendje) jo direct dukshem.Ky process themelor mund te jete lidhur vetem probabilisht me nje process tjeter te dukshem stokastik duke prodhuar sekuencen e tipareve apo karakteristikave qe ne mund te vezhgojme.Nje model i fshehur i Markovit eshte nje zinxhire Markovi ku output i vrojtuar eshte nje variable X i gjeneruar sipas nje funksioni probabilitar output qe i shoqerohet cdo gjendjeje.

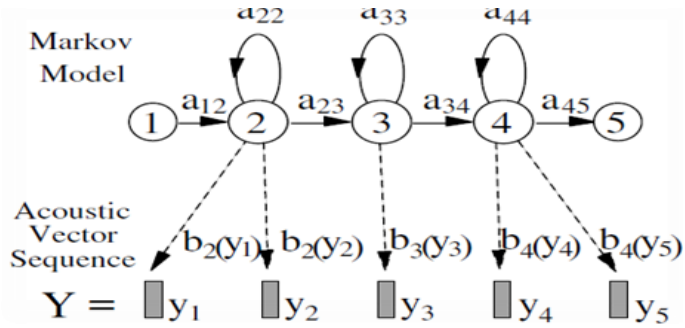


Figura 3.5 : Modeli telefonik me baze HMM me 5 gjendje

Formalisht nje model i fshehur i Markovit mund te percaktohet si me poshte:

- $Y = \{y_1, y_2, \dots, y_M\}$ - nje alphabet vrojtimi output .Simbolet e vrojtimit korespondojne me outputin fizik te sistemit te modeluar.
- Ω=$\{1,2,..,N\}$- nje grup gjendjesh qe perfaqesojne hapesiren e gjendjes.
- $A = \{a_{ij}\}$- nje matrice kalimi i probabiliteteve ,ku a_{ij} eshte probabiliteti i marrjes se nje kalimi nga nje gjendje *i* ne nje gjendje *j*.

  $a_{ij} = P(s_t = j | s_{t-1} = i)$
- $B = \{b_i(k)\}$-nje matrice probabilitare output,ku $b_i(k)$ eshte probabiliteti i emetimit te simbolit y_k kur gjendja e *i*-te shenohet.Le te jete $X = X_1, X_2, \dots, X_t, ..$ outputi i vrojtuar i HMM.Sekuenca e gjendjeve $S = s_1, s_2, \dots, s_t, \dots$ nuk eshte vrojtuar (fshehur) dhe $b_i(k)$ mund te rishkruhen si me poshte :

  $$b_i(k) = P(X_t = y_k | s_t = i)$$
- $\pi = \{\pi_i\}$- nje shperndarje e gjendjes fillestare ku :

  $$\pi_i = P(s_0 = i) 1 \le i \le N$$

Meqenese $a_{ij}, b_i(k), dhe\ \pi_i$ jane probabilitete ,ato duhet te kenaqin keto veti :

$a_{ij} \geq 0, b_i(k) \geq 0, \pi_i \geq 0 \quad \forall\, all\ i,j,k$

$$\sum_{j=1}^{N} a_{ij} = 1$$

$$\sum_{k=1}^{M} b_i(k) = 1$$

$$\sum_{i=1}^{N} \pi_i = 1$$

Parametrat e modelit akustik $= \{a_{ij}, b_i(k)\}$, vleresohen ne menyre efikase nga nje korpus i shprehjeve te trajnimit duke perdorur lgoritmin perpara-prapa.Ne perfundim pershkrimi i plote i nje HMM perfshin parametra ted y madhesive N dhe M ,qe perfaqesojne numrin e gjendjeve dhe madhesine e alfabeteve te vrojtuar , alfabetin e vrojtuar Y dhe tre matrica te masave probabilitare A,B dhe π. Shenimi I meposhtem :

$$\Phi = (A, B, \pi)$$

Perdoret per te treguar gjithe bashkesine e parametrave ten je HMM.

Duke patur parasysh perkufizimin e mesiperm te HMMs ,mund te formulojme tre problem themelore :

A. Problemi i vlefshmerise : Jepet nje model Φ dhe nje grup vrojtimesh $Y = (Y_1, Y_2, \dots, Y_T)$,kush eshte probabiliteti $P(Y|\Phi)$ qe ky grup vrojtimesh Y te jete gjeneruar nga modeli Φ?
B. Problemi i dekodimit : Jepet nje model Φ dhe nje grup vrojtimesh $Y = (Y_1, Y_2, \dots, Y_T)$,kush eshte grupi apo vargu me i mundshem i gjendjeve $S = (s_0, s_1, \dots, s_T)$ ne model qe prodhon vrojtimet ?I
C. Problemi i mesimit : Jepet nje model Φ dhe nje grup vrojtimesh,si mund te rregullojme parametrat e modelit per te maksimizuar shanset e perbashketa (mundesia) $\prod_x P(X|\Phi)$?

Me zgjidhjen e problemit te vleresimit ne jemi ne gjendje qe te vleresojme sa me mire nje HMM te dhene ne perputhje me nje grup te dhene vrojtimi.Prandaj ,HMM eshte perdorur per te bere njohjen e modelit ,pasi pergjasia $P(X|\Phi)$ eshte perdorur per te llogaritur probabilitetet e pasme (posterore) $P(\Phi|X)$,dhe HMM me probabilitetin me te larte posterior percaktohet si modeli i deshiruar per grupin e vrojtimit.Me zgjidhjen e problemit deshifrim ne mund te gjejme grupin me te mire te gjendjeve qe perputhen per nje grup vrojtimi te dhene , ose me fjale te tjera ne mund te zbulojme grupin e fshehur te gjendjes.Dhe me zgjidhjen e problemit te te mesuarit ,ne do te kemi mjete per te vleresuar automatikisht parametrin Φ te modelit nga grupi i trajnimit.Detyra me e veshtire eshte nje mesim ,meqe nga te dhenat e trajnimit duhet te vleresoje parametrat e HMM te tille qe te mund te karakterizojne njesine e zgjedhur te te folurit.Sinjali vocal eshte i ndare ne njesi elementare si : fjale ,fonema,tri-fonema.Per cdo njesi i shoqerohet nje HMM ,dhe per cdo gjendje ten je HMM nje dritare kohore,me parametra zanore te llogaritur per kete dritare specifike.Gjate fjales trakti vocal kalon neper nje sekuenve apo grup te gjendjeve (qe modelohen me gjendjet HMM) dhe ne cdo gjendje nje segment nga fjalimi vocal eshte emetuar nga me nje vector te parametrave qe perbejne prodhimin apo outputin e gjendjeve te HMM .Ky vector output e HMM duhet te marre vlera te vazhdueshme pasi parametrat zanore marrin vlera ne nje hapesire te vazhdueshme.Per kete arsye perzierjet Gausiane jane perdorur per te modeluar vezhgimet apo vrojtimet e gjendjeve te HMM .Cdo parameter i vektorit output mund te modelohet me nje shume peshe te funksioneve me shperndarje normale:

$$b_j(y_t) = \sum_{g=1}^{G} w_{jg} N(y_t, \mu_{jg}, \Sigma_{jg})$$

Ku :

$$N(Y, \mu_i, \sum_i) = \frac{1}{\sqrt{(2\pi)^n |\sum|}} e^{-\frac{1}{2}(Y-\mu_i)^T \sum^{-1} (Y-\mu_i)}$$

Ku : $Y = [y_1, y_2, \dots, y_n]$ eshte nje vector vrojtimi n-dimensional ,n eshte numri i parametrave te zerit te cilat jane nxjerre nga vrojtimi . $|\sum|$ eshte percaktori i matrices diagonale te kovariances ,G eshte numri i komponenteve apo perberesve te perzierjes,dhe

w_{jg} eshte pasha e komponentit g te gjendjes j.Modelimi i fjales duke perdorur modelet e fshehura te Markovit bend dy supozime :

- Procesi i Markovit : sekuenca apo grupi gjendje ne nje HMM eshte supozuar te jete nje proces Markovi i rendit te pare ,ne te cilin probabiliteti i kalimit ne gjendjen tjeter varet vetem nga gjendja aktuale ,keshtu qe historia e gjendjeve te meparshme nuk eshte e nevojshme .
- Pavaresia e vrojtimit :Vrojtimet jane te pavarura me kusht nga te gjithe vrojtimet e tjera qe japin gjendjen qe gjenerojne ate.

Keto dy supozime mund te cojne ne nje model jorel te fjales ,por jane te nevojshme per shkak te thjeshtezimeve matematike dhe llogaritjeve qe sjellin .Problemi i vleresimit dhe dekodimit do te jene shume te veshtira per tu trajtuar pa keto supozime.Megjithate dy dekadat e fundit te suksesit te HMMs ne modelimin e sinjalit te fjales kane provuar se keto supozime nuk jane te rendesishme .

3.7 Nxjerrja e tipareve

Meqenese modelet HMM nuk modelojne direct formen valore te sinjalit akustik,ne kete pjese do te diskutohet nje lloj perpunimi akustik qe zakonisht quhet nxjerrja e tipareve ose analiza e sinjalit ne literature e njohjes se fjales.Termi karakteristike apo tipar I referohet vektorit te numrave qe paraqet nje kohe-pjese te sinjalit te fjales.Llojet me te perdorura te karakteristikave apo tipareve jane karakteristika LPC ,karakteristika PLP dhe ajo MFCC.Keto jane quajtur karakteristika spektrale sepse ato paraqesin formenvalore ne drejtim te shperndarjes se frekuencave te ndryshme qe perbejne formenvalore.

Parametrat e fjales perpunohen shpesh nga filtrat.Filtrimi me I zakonshem ndodh ne nivelin spectral,ku fuqia e spektrit eshte perpunuar permes kanaleve te bankave te filtrit.Karakteristika MFCC perdor banken e filtrave te shkalles Mel nderkohe qe karakteristika PLP perdor shkallen Bark per analizen e zonave te saj kritike.

Karakteristika e pare qe perdorim eshte vete formavalore e fjales.Fakti qe njerezit ,dhe disa makina ekzistuese ,jane te afte per transkriptimin dhe te kuptuarin e fjales vetem duke dhene valen e zerit con ne perfundimin se formavalore permban informacion te mjaftueshem per te

kryer kete detyre.Ndonjehere ky informacion eshte i veshtire qe te zberthehet vetem duke u perqendruar ne formenvalore ,por dhe nqse nje inspektim visual eshte i mjaftueshem per te terhequr disa karakteristika te rendesishme.Per shembull dallimi midis zanoreve dhe disa bashketingelloreve eshte relativisht i qarte ne nje formevalore .Zanoret karakterizohen nga nje konfiguracion i hapur i traktit vocal.Ky contrast me bashketingelloret ,te cilat karakterizohen nga nje shtrengim apo mbyllje ten je ose me shume pika pergjate traktit vocal.Kjo perkthehet ne nje diference te dukshme te energjise .Hulumtuesit kane aftesine qe te shohin ne spektogram dhe te identifikojne disa zanore ose bashketingellore per shkak te amplitudes se tyre.Ne figuren e meposhtme eshte nje paraqitje e energjise se zanoreve (“a”, “e” ,”i” ,”o” ,”u”) e sigururar nga paketa Matlab .Mes tyre, zonat ku energjia eshte pothuajse zero,jane momentet e heshtjes ,kur folesi pushon para se te levizi per ne zanore tjeter.

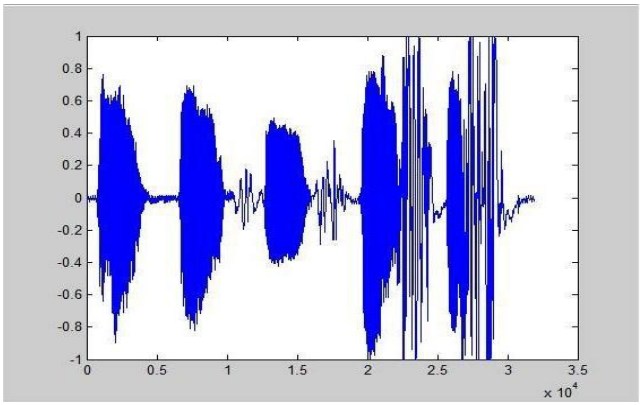
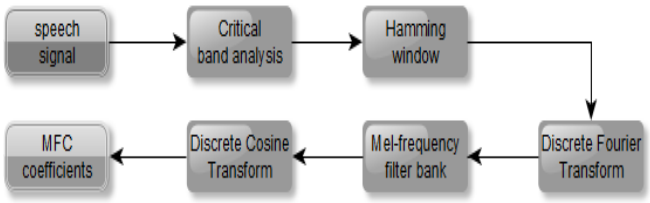


Figura 3.6 : Ilustrimi i amplitudës së zanoreve

Kur sinjali eshte stacionar (i palevizshem) ,analiza spektrale nuk mund te kryhet ne te gjithe sinjalin ,por ne korniza te shkurtra (20-30ms) ne te cilen sinjali eshte kuazi-stacionar.Sinjali original eshte segmentuar ne domenin kohor,duke perdorur nje dritare Hamming,dhe procesi i nxjerrjes se tipareve eshte kryer ne cdo dritare te vetme.Ne pergjithesi ,karakteristikat e domenit kohor jane shume me pak te sakta se sa karakteristikat e domenit frekuencor te tille si koeficienti cepstral i frekuencave Mel (MFCC).Kjo thuhet per shkak se shume karakteristika te tilla si formantet ,te perdorura ne diskriminimin e zanoreve,jane te karakterizuara me mire ne domenin frekuencor .Kur llogarisim koeficientet MFCC ,eshte perdorur nje shkalle e frekuencave jo-lineare,pasi perafron me mire sistemin e degjimit te njeriut .Ky process analizimi merr parasysh se kapja e toneve te ndryshme te zerit eshte bere

ne nje shkalle logaritmike brenda veshit ,ne menyre propocionale me frekuencen themelore te zerit .Ne kete menyre ,reagimi i veshit te njeriut eshte jo-linear ne lidhje me frekuencen ,pasi eshte ne gjendje te ndjeje ndryshime te vogla te frekuencave mes komponenteve me frekuence te ulet me lehte se sa te atyre me frekuence te larte .Per te percaktuar koeficientet cepstral,spektri,i llogaritur duke perdorur FFT ,eshte me i zbutur neper disa banka filtrash trekendesh ,cdo qenderzim ne nje frekuence gjendet ne shkallen Mel.Shkalla Mel eshte nje shkalle perceptual e ndertimit te regjistrave duke u nisur nga disa degjues te cilet jane te barabarte ne distance nga njeri-tjetri.Qellimi i ketij grupi te filtrave trekendore eshte ajo e ndarjes se sinjalit mbi zonat e frekuencave te lidhura me shkalle Mel.Per nje sinjal me ze me nje gjeresi prej 8kHz ,nje numer prej 24 grupe filtrash konsiderohet i mjaftueshem per te llogaritur parametrat MFCC.Megjithate ne sistemet e njohjes se fjales ky numer eshte konfiguruar, dhe permes eksperimenteve mund te gjendet vlera e saj optimale per aplikimin perkates.Duke aplikuar ngjeshjen logaritmike ne outputin e bashkesise se filtrave ,shperndarja e koeficienteve ndjek nje ligj Gausian.Pastaj mbi cdo band eshte llogaritur energjia mesatare.Koeficientet MFCC jane marre pas aplikimit te transformimit diskret kosinus DCF,i cili eshte nje instrument shume i pershtatshem .Ai merret vetem me numrat real,ka nje veti te forte te ‘’kompaktesimit te energjise’’ .Shumica e informacionit te sinjalit tenton te perqendrohet ne nje component me frekuence te ulet dhe zberthen keto vlera.



Me fjale te tjera hapat e nxjerrjes se koeficienteve MFCC:

- Ne fillim llogaritet transformimi Furie i nje sinjali (nje fragment i dritarizuar):

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-2j\pi nk}{N}} \qquad 0 \leq k \leq N$$

- Bashkesia apo grupi i M (m=1,2,…,M) dritareve trekendeshe te mbivendosura eshte percaktuar ne menyre qe te ndaje kopetencat e spektrit te marre me siper ne shkallen Mel:

$$H_m[k] = \begin{cases} 0, & k < f[m-1] ose\ k > f[m+1] \\ \dfrac{2(k - f[m-1])}{(f[m] - f[m-1])(f[m+1] - f[m])}, & f[m-1] < k < f[m] \\ \dfrac{(f[m+1] - k)}{(f[m+1] - f[m])(f[m] - f[m-1])}, & f[m] < k < f[m+1] \end{cases}$$

Formula gjithashtu mund te shprehet dhe si me poshte :

$$H_m[k] = \begin{cases} 0, & k < f[m-1] ose\ k > f[m+1] \\ \dfrac{(k - f[m-1])}{(f[m] - f[m-1])}, & f[m-1] < k < f[m] \\ \dfrac{(f[m+1] - k)}{(f[m+1] - f[m])}, & f[m] < k < f[m+1] \end{cases}$$

Ne kete rast $\sum_{m-1}^{M} H_m[k] = 1$. E vetmja qe ndryshom midis dy perfaqesive te mesiperme eshte nje vector i konstanteve per te gjithe sinjalet e dhena input,aq sa I njejti filter eshte perdorur kudo,zgjedhja e te cilit eshte e parendesishme.

Ky grup i fitrave llogarit spektrin rreth frekuencave qendrore te cdo bande (grupi).Bandat e tyre rriten sipas indeksit m.

Banka e pare e filtrit do te filloje ne piken e pare ,arrin kulmin e saj ne piken e dyte ,pastaj kthehet ne zero ne piken e trete.Banka e dyte e filtrit ,do te filloje ne piken e dyte ,,do te arrije kulmin e saj ne piken e trete,dhe do te behet zero ne piken e katert.Ploti perfundimtar I mbulimit te M filtrave duket si me poshte:

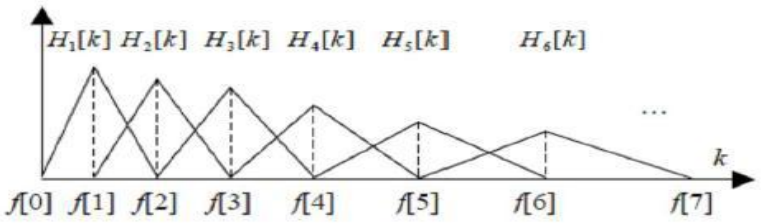


Figura 3.7 : Bandat e frekuencave ne shkallen Mel

Le te konsiderojme qe: f_1, f_h jane me te ultat,perkatesisht frekuenca me e larte ne banken e filtrave,e shprehur ne Hz ,tjetra eshte frekuenca e modulimit apo mostrimit e shprehur gjithashtu ne Hz, M numri i filtrave dhe N madhesia e dritares FFT.Pikat e kufirit jane vendosur ne menyre te njetrajtshme neper shkallen Mel:

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}(B(F) + \frac{B(f_h) - B(f_l)}{M+1})$$

Ku shkalla Mel B eshte:

$$B(f) = 1125\ln(1 + \frac{f}{100})$$

Dhe i anasjellti B^{-1} eshte :

$$f = 700(\exp\left(\frac{b}{1125}\right) - 1)$$

- Pastaj, merren logot e fuqive ne cdo frekuence mel(vakt).

$$S[m] = \ln\left[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right], \qquad 0 \le m \le M$$

- DCT eshte aplikuar ne listen e M log fuqive Mel ,sikur te ishte nje sinjal

$$c[n] = \sum_{m=0}^{M-1} S[m]\cos(\frac{n\pi(m+1/2)}{M}), \ \ 0 \le n \le \mathrm{M}$$

- Koeficientet MFCC jane amplitudat e spektrit resultant.Vetem 2-13 koeficiente DCT jane marre ,pjesa tjeter eshte fshire.

Ndryshimet e perkohshme ne spectra luajne nje rol te rendesishem ne perceptimin njerezor.Edhe pse eshte llogaritur secili prej grupeve te koeficienteve mbi nje dritare te shkurter Hamming,informatat qe gjinden nga dinamika kohore e ketyre parametrave eshte shume i dobishem ne njohjen automatike te te folurit.Nje menyre per te kapur kete informacion eshte duke perdorur koeficientet delta,,qe masin ndryshimin e koeficienteve me kalimin e kohes .Ata jane te njohur dhe si koeficiente diferencial dhe pershpejtimi.Del se llogaritje e trajektoreve MFCC dhe bashkengjitja e tyre vector funksionit fillestar do te rrise ndjeshem performance e sistemit automatic te njohjes se fjales .Informacioni kohor eshte vecanerisht plotesues per HMM ,pasi modelet HMM supozohen te pavarura nga e kaluara,ne dallim me keto karakteristike dinamike qe zgjerojne fushen e kornizes(frames).Kur shkalla e

mostrimit eshte 16-kHz ,nje system tipik i fjales state-of-arte mund te ndertohet duke u bazuar ne karakteristikat e meposhtme:

_ rendi i 13 I koeficienteve MFCC c_k

_ rendi i 13 I 40-msec delta koeficientet MFCC te rendit te pare llogariten nga

$$\Delta c_k = c_{k+2} - c_{k-2}$$

_ rendi i 13 I 40-msec te delta koeficienteve MFCC te rendit te dyte llogariten nga

$\Delta\Delta c_k = c_{k+1} - c_{k-1}$

Analizimi me kohe-shkurter me 256 ms dritare Hamming eshte perdorur zakonisht per te llogaritur MFCC. $c_k[0]$ eshte perfshire ne vector funksionin.Ne perfundim ,vector funksioni i perdorur per njohjen e fjales ,eshte ne pergjithesi nje kombinim i ketyre tipareve:

$$x_k = \begin{pmatrix} c_k \\ \Delta c_k \\ \Delta\Delta c_k \end{pmatrix}$$

Dhe eshte provuar qe merren rezultate shume te mira.Ai eshte formuar nga 39 koeficiente: 12 koeficiente MFCC + energjine,se bashku me derivatet kohore te tyre te rendit te pare dhe te dyte.

3.8 Zgjedhja e njësisë themelore

Detyra e zgjedhjes se nje njesie te duhur nuk eshte aq e thjeshte sa mund te duket.Ne menyre qe te hartojme nje system te realizueshem ,duhet te merren ne considerate disa ceshtje te rendesishme kur zgjedhim me shume njesi baze.

- Njesia duhet te jete e sakte ,per te paraqitur realizimin akustik qe shfaqet ne kontekste te ndryshme.
- Njesia duhet te jete e trajnuar .Per te vleresuar parametrat e njesise ,duhet te kete te dhena te mjaftueshme ne dispozicion.Ketu vihet ne dukje dhe nje here pse fjalet jane zgjedhje me pak e trajnuar ne ndertimin e nje sistemi te njohjes,pasi qe ,pavaresisht saktesise se tyre ,eshte pothuajse e pamundur per te marre disa qindra perseritje per te gjitha fjalet .Fjalet jane nje zgjedhje e duhur per njesine baze vetem ne rastin kur njohja e fjales eshte domein specific dmth vetem per shifra.
- Njesia duhet te jete e gjenerueshme,ne menyre qe cdo fjale e rete nxirret nga nje njesi e paracaktuar e ruajtur per njohjen e fjales .Nese ky record do te konsistonte ne nje

grup te caktuar te modeleve te fjales ,nuk mund te kete ndonje menyre per te nxjerre modelin e ri te fjales.

Nje sfide praktike eshte se sit e balancojme keto tre kritere te rendesishme.Keshtu ,ne vend te fjaleve te modelimit ,sistemet e njohjes se fjaloreve-te medhenj ,perdorin nen-fjalet si njesi baze te te folurit,per shembull fonet ku fjalet nuk jane as trajnuar as gjeneruar.

Foni *është një tingull fjalim ,segmenti diskret me i vogël i tingullit ne nje rrjedhe te fjales.*

Modelet fonetike nuk japin problem trajnimi, meqenesedukuri te mjaftueshme per te gjithe fonet mund te gjenden ne disa mije fraza,atom und te jene trajnuar ne nje detyre dhe testuar ne nje tjeter sepse ato nuk varen nga fjalori.Keto i bejne fonet te trajnueshem dhe te gjeneralizuar.Megjithate ky model fonetik supozon se nje foneme eshte identike ne cdo kontekst .Ky nuk eshte rasti i vetem sepse fonemat nuk jane prodhuar ne menyre te pavarur dhe realizimi i nje foneme varet fuqimisht nga fonemat fqinje.Mund te themi qe keto modele fonetike cojne ne modele me pak te sakta.

Kjo pengese mund te tejkalohet nese kemi parasysh kontekstin e njesive te varura.Ne qofte se ne kemi nje trajnim te madh te mjaftueshem te vendosur per te vleresuar keto parametra ne kontekst-te varur,ne mund te permiresojme ndjeshem saktesine e njohjes.Kemi prezantuar idene e modelimit tri-fone,nje model fonetik qe merr ne considerate ted y fonet fqinje te majte dhe te djathte . Nese dy fonet kane te njejtin identitet por kontekst te majte apo te djathte te ndryshem ,ato konsiderohen si trifone te ndryshme .Realizimet e ndryshme ten je foneme jane percaktuar me termin allofon.

Modelet tri-fone jane modele shume te fuqishme fonetike dhe jane me te qendrueshme se sa njesite ne kontekstin e pavarur ,por ne kete rast ,trajnimi behet nje detyre sfiduese.Meqenese cdo kontekst tri-fone eshte i ndryshem ,idea kryesore eshte qe te gjenden rastet e konteksteve te ngjashme dhe ti bashkojme ato ne menyre qe te meret nje numer i menaxhueshem modelesh qe mund te trajnohen mire.

Duke shkuar nje hap me tej ,me qellim te balancimit te trajnueshmerise dhe sigurise midis modeleve fonetike dhe fjales ,verehet modelimi i ngjarjeve nen-fonetike .Per modelimin nen-fonetike ,ne mund te trajtojme apo perdorim gjendjen ne modelet HMM si njesi baze nen-fonetik. Ne kete kontekst ,koncepti i modeleve te fshehura te Markovit eshte propozuar dhe pergjithesuar per shperndarjet output te gjendjes-varur pergjate modeleve te ndryshme fonetike.

3.9 Vlefshmëria (Saktësia) e ASR

Per nje kontroll sa me te mire te sistemit ASR ,eshte mire qe te perdoret nje moster e vogel nga te dhenat e trajnimit per te matur performance e caktuar te trajnimit.Performanca e trajnimit eshte e dobishme ne fazen e zhvillimit per te identifikuar te metat e mundshme te zbatimit.Perfundimisht ,testet duhet te behen ne nje grup te zhvillimit qe zakonisht perbehet nga te dhena qe nuk jane perdorur kurre ne trajnim.

Nje menyre per te vleresuar performance e modelit gjuhesor eshte te vleresojme shkallen e gabimit te fjales (WER) te dhene ,kur vendors ne nje system te njohjes.WER eshte nje mjet shume i pershtatshem qe perdoret kur krahasohen modele te ndryshme gjuhesore ,si dhe per vleresimin e permiresimeve brenda nje sistemi. Veshtiresia me e madhe kur perdorim kete metode qendron ne faktin qe sekuenca e njohjes se fjales mund te kene nje gjatesi te ndryshme nga sekuenca e referimit ,i cili duhet te jete i sakte.Keto dy sekuenca fjalesh jane te lidhura me nje algoritem qe ka si synim final minimizimin e kostos se redaktimit te shprehjes se njohjes,ne menyre qe te duket njelloj me ate referim.WER mund te llogaritet si me poshte :

$$WER[\%] = \frac{\#shtime + \#zevendesime + \#fshirje}{\#Numri\ i\ fjaleve\ ne\ kopjimin\ e\ referimit} X\ 100$$

Zakonisht jane tre lloje te gabimeve te njohjes se fjales ne njohjen e te folurit:

- Zevendesimi :Nje fjale e gabuar zevendesohet nga fjala e sakte
- Fshirja : Nje fjale e sakte ishte lene pas dore ne fjaline e njohur
- Shtimi : Nje fjale shtese u shtua ne fjaline e njohur.

Ky lloj vleresimi i performances megjithate nuk jep detaje specifike ne lidhje me natyren e gabimeve te perkthimit dhe puna ne vijim kerkon te gjendet burimi i gabimit.Per me teper ,ky lloj i matjes nuk permban faktin qe gabim zevendesimi mund te hiqet apo te fshihet lehte nese numri i karaktereve te gabuara eshte i vogel (si psh fjalet “ruajtur” dhe “luajtur”), apo veshtire te hiqen nese numri i karaktereve te gabuara do te jete i madh (psh “ruajtur “ dhe “mbeshtetur”). Ne menyre qe te kapercehet kjo pengese ndonjehere perdoret nje vleresues ne nivel karakteri:

$$CHER[\%] = \frac{\#Ch.shtime + \#Ch.zevendesime + \#Ch.fshirje}{\#Numri\ i\ karaktereve\ ne\ kopjimin\ e\ referimit} X\ 100$$

Metoda e fundit e vleresimit behet ne nivelin e fjalive,dhe eshte e dobishme vetem ne rastet kur transkriptimi i gabuar i nje fjale te vetme ne nje sekuence fjalesh e ben njohjen te pavlefshme.

$$SER[\%] = \frac{\#Jo\ fjali\ e\ gabuar}{\#Fjalite\ ne\ kopjimin\ e\ referimit} X\ 100$$

Nga keto tri karaktere te performances te perdorura ne vleresimin e nje sistemi te njohjes,me e perdorur zakonisht eshte WER.

3.10 Mjetet e njohjes së fjalës

Projekti CMU Sphinx është një prej projekteve më të vjetra Open Source e cila tenton të sjellë problemin e njohjes së gjuhës së folur (Njohjen e Ligjëratës) pranëzhvilluesit dhe kërkuesit. Sipas informacionit të publikuar nga grupi i punë së Sphinx-4, thuhet se Sphinx-4: "është dizenjuar së bashku nga CMU Mellon University,laboratorët e Sun Microsystems, Mitsubishi Electric Research Lab, dhe Hewlett –Packard Cambridge Research Lab" dhe vazhdon më pas duke shpjeguar se arsyeja e tëshkruarit në Java është që ta bëjë "tepër të lëvizshme".

Pra në këtë punim është përdorur ky system ,CMU Sphinx, sepse integron me sukses metodën statistikore të modeleve të fshehura të markovit dhe gjithashtu mund të trajnojë dhe të realizojë modele fonetike në një rrjet të sofistikuar leksikor deshifrimi . Ky është një mjet shumë popullor dhe përdoret zakonisht njohjen e të folurit, sepse ofron mundësinë e zhvillimit të gjuhëve të pavarur(siç është gjuha shqipe) dhe gjithashtu paraqet përmbledhjen e shumë fjalëve që zëvëndësohen/fshihen /futen dhe mund të llogarisë normën e gabimit të fjalisë/fjalës në një mënyrë të përshtatshme .

3.11 Përgatitja e bazës akustike

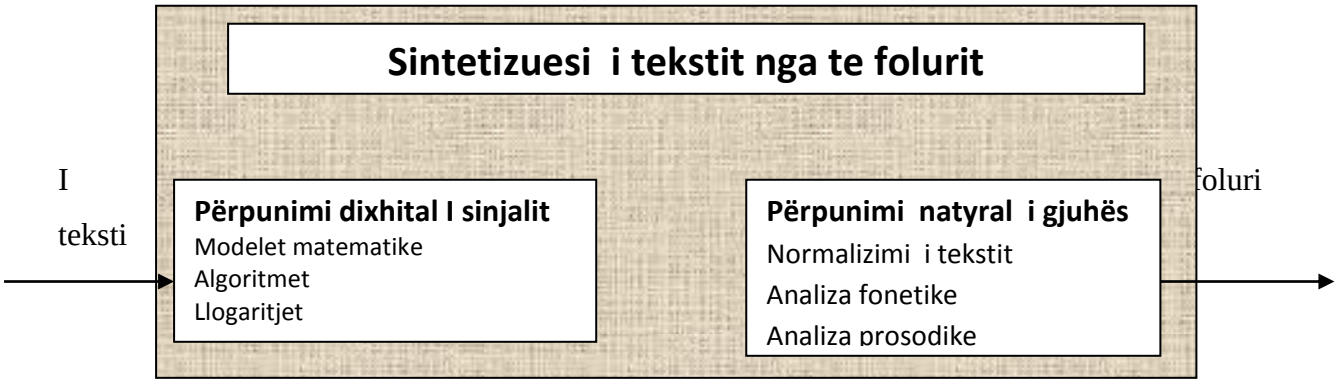


Figura 3.8 *Skema e pergjithsjme per shnderrimin e te folurit ne tekst*

Kjo skemë paraqet platformën bazë për ndërtimin e sistemeve ASR dhe tanimë është përvetësuar nga shumica dërmuese e autorëve si një postulat i qëndrueshëm. Në anën tjetër, mënyra e realizimit të funksioneve të këtyre moduleve mbetet çështje e hapur dhe me shumë ide.

Parapërgatitja e strukturës së një sistemi ASR, si hap të parë dhe të pashmangshëm nënkupton dizajnimin e burimit, prej nga sistemi do të ushqehet me të dhëna akustike. Të dhënat akustikemund të jenë të llojeve të ndryshme dhe të ruajtura në forma të ndryshme.

Për të ndërtuar një model të njohjes së fjalës shqipe ,meqë nuk ekziston një bazë konkrete të dhënash apo korpus i ndërtuar për aplikime atëhere një bazë të vogël të dhënash e ndërtuam duke u nisur nga segmente akustike të përfruara apo të nxjerra nga një lajm në internet. Një nga arsyet e zgjedhjes së këtyre lloj artikujve është disponueshmëria e të dhënave tekst dhe të folurit në të njëjtën kohë.

- Të dhënat e të folurit dhe modeli akustik

Lloji dhe sasia e të dhënave të të folurit ndryshojnë në varësi të gjuhësh që shqyrtohet.Të dhënat që kemi marrë në shqyrtim për gjuhën shqipe përbëhen nga trasmetimi i një lajmi konkret i marrë në internet i lexuar nga një folës i vetëm mashkull (F1) .Këto të dhëna janë ndarë në grupe trajnimi dhe testimi.Nga këto 67 fjalë janë për testim dhe 68 fjalë janë për trajnim.

Pra kemi përzgjedhur sinjalin në mënyrë periodike nga një regjistrim i brëndshëm kompjuterik ,ku kemi kaluar një lajm të regjistruar në youtube në një moment të caktuar ,si një sekuencë kampjonësh të gatshëm për tu dëgjuar.

Formati më i thjeshtë i ruajtjes së të dhënave është i ashtuquajturi skedari WAV. Në këtë skedar ruhen si vlera reale amplitudat e sinjalit gjatë momentit të kampionimit(përzgjedhjes). Paraqitja e informacionit në skedar bëhet duke e ruajtur informacionin kampion për kampion deri në fund të gjithë regjistrimit.Arsyeja e zgjedhjes së këtij skedari në rastin e simulimit lidhet me thjeshtësinë që e karakterizon këtë tip skedari. Këtu informacioni ruhet në mënyrë të drejtpërdrejt, pa ndonjë lloj kompresimi apo modifikimi. Ky format skedari nuk është me pagesa pra prodhimi apo procesimi i tij mund të bëhet pa ndonjë liçensë të veçantë.Më pas skedari lexohet nga një lexues parametrik i skedarit ,i cili copëzon informacionin e gjendur në file-in WAV dhe e kalon më pas në copa prej 1024 kampionësh për të prodhuar koefiçientët MFCC. Vlen të përmendet në këtë moment se në rastin kur korniza e fundit ka më pak se 1024 kampionë të mbetur, lexuesi e mbush (padding) kornizën me vlera 0 për të dhënë rezultat vektorin me gjatësi sa fuqia e 2-shit e cila bën të mundur përdorimin e IDFT për një procesim më të shpejtë sinjalit.

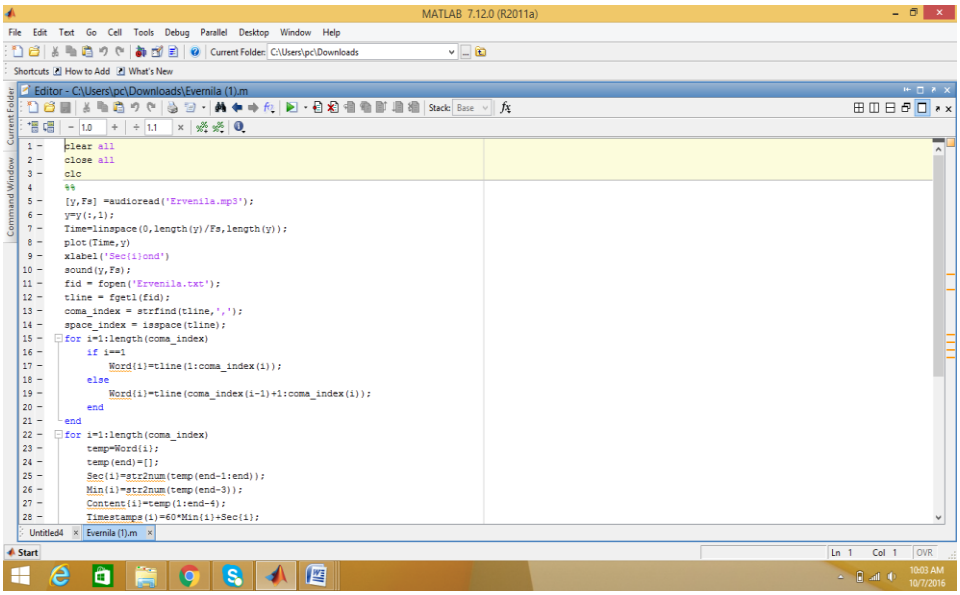


Figura 3.9 *: Dokumenti korpus(programi) i nxjerrjes së lajmit(databazës)*

Kjo databazë konsiston në përafërsisht 2 minuta apo 120 sekonda (50 shprehje) të një folësi të vetëm mashkull për leximin e lajmit konkret *" Ekskluzive/BE "Njeh" Çështjen Çame –Top Channel Albania-News*, datë 30.07.2016.

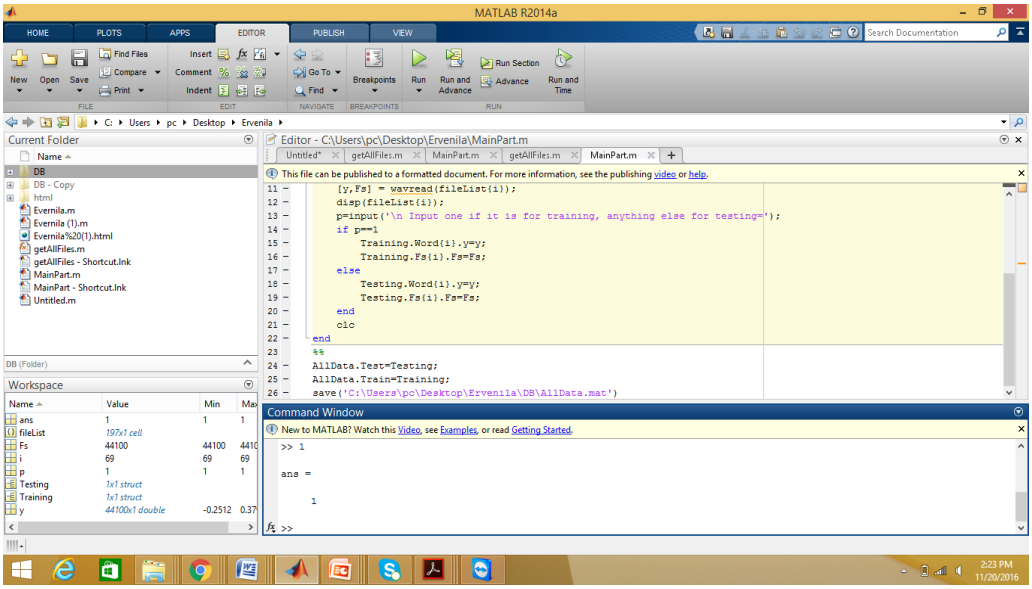


Fig 3.9 Pamje e krijimit të testim dhe trainim database

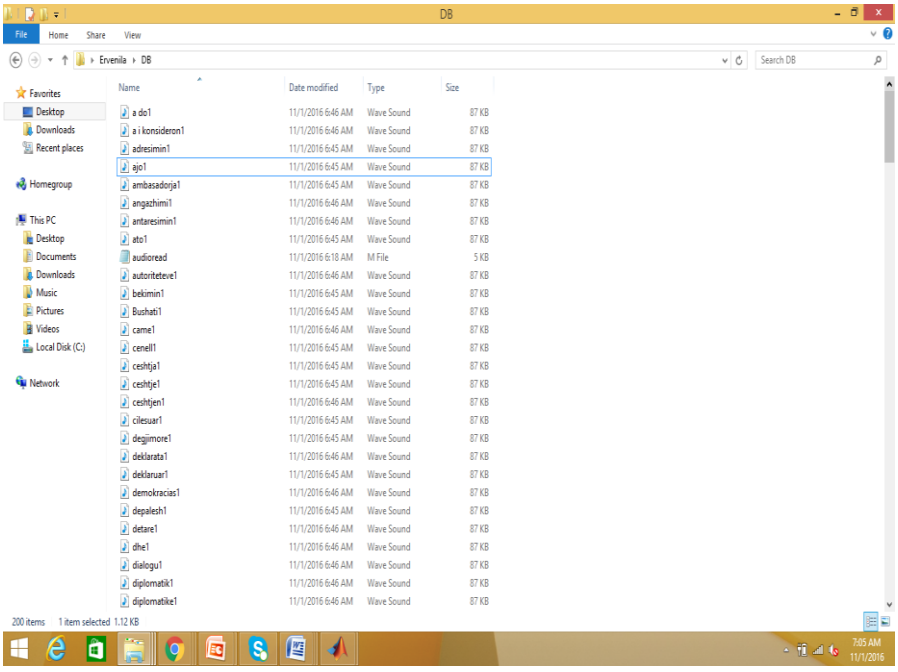


Fig 3.10 : Pamje e krijimit të databazës

KAPITULLI IV

Optimizimi i njohjes së fjalës duke përdorur algoritmin PSO (particle swarm optimization)

Në fushen e njohjes së fjalës ,studiuesit kane propozuar lloje te ndryshme te metodave te optimizimit ne te dy skajet.Edhe pse ka shume algoritme te suksesshme si Baum-Welch dhe metoda gradient te cilat jane zhvilluar per optimizimin e parametrave Pengese e ketyre metodave eshte se ato jane algoritme koder-ngjitje dhe varen mjaft fuqishem ne vleresimet fillestare te parametrave te modelit .Cdo vleresim arbitrar I parametrave fillestare te modelit zakonisht do te coje ne nje model nen-opimal ne praktike .

Nje nga menyrat me efektive per te permiresuar fuqine e njohjes së fjalës, eshte gjetja e tipareve apo karakteristikave invariante (te pandryshueshme) ose te fuqishme ne menyre qe te minimizohet ndryshueshmeria e vrojtimit te shkaktuar nga lloje te ndryshme faktoresh dhe mosperputhjesh midis trajnimit dhe testimit te kushteve.

Ndoshta tiparet me te njohura sot dhe te perdorura per njohjen e te folurit jane koeficientet cepstral te frekuencave Mel (MFCC) te cilat jane marre fillimisht nga transformimi Furie me kohe te shkurter (STFT) dhe pastaj analiza e bankave te filtrave ne spektrin e frekuencave te shkalles Mel.Ka shume ceshtje me keto karakteristika .Se pari ato jane shume te ndjeshme ndaj zhurmes se ambjentit dhe performance e tyre gjithashtu degradon kur kushtet e testimit dhe te trajnimit nuk jane te njejta.Se dyti brezi I filtrave dhe numri I filtrave nuk jane te standartizuara.Normalisht 20 deri 40 filtra trekendesh jane perdorur per te zbatuar MFCC te cilat jane ne hapesiren lineare para 1 kHz dhe ne hapesiren logaritmike pas 1 kHz.Per te zgjidhur kete ceshtje ky kapitull propozon nje perafrim me perdorimin e teknikes se optimizimit te tufes se grimcave (PSO) per te optimizuar koeficientet MFCC.Me ndihmen e ketij algoritmi numri dhe ndarja e filtrave jane optimizuar ne kushte normale ne terren dhe ne ambjent te zhurmshem .Shumica e studimeve per te otimizuar njohjen e fjalës, eshte fokusuar ne prapa-fund dhe ka shume pak hulumtime ne perdorimin e ketyre teknikave ne para-fund .

Sic do te shohim ne kete kapitull e reja e punes qendron ne :

1) Nxjerrjen e bankes se filtrave Mel me ndihmen e optimizimit te tufes se grimcave
2) Optimizimi I peshuar I MLPs(multi layer perceptron) duke perdorur PSO

4.1 Teknika optimizimi te nxitura nga natyra

Sistemet natyrale jane nje nga burimet me te pasura te frymezimit per shpikjen e sistemeve te reja inteligjente .Me analizen e sistemeve natyrore jane shpikur nga studiuesit mjete dhe teknika te reja te rendesishme gjate dy dekadave te fundit sic jane algoritmi gjenetik ,GA, optimizimi i grimcave,PSO,optimizimi i kolonisa se milngonave dhe dinamika e formimit te lumit .Midis tyre

si me premtuese kane dale algoritmet GA dhe PSO per shkak te shkathtesise se tyre dhe aftesise per te optimizuar ne hapesira komplekse multimodale aplikuar ne funksionet kosto jo-te ndryshueshem.Te dyja keto teknika jane te ngjahsme ne kuptimin qe keto teknika jane metoda kerkimi me baze popullsine dhe kerkojne nje zgjidhje optimal nga perditesimi i brezave .Keto dy teknika jane menduar per gjetjen e zgjidhjes per nje funksion te caktuar objektiv ,por perdorin strategji te ndryshme dhe aftesi kompjuterike te ndryshme.

4.2 Optimizimi i pjesshëm i tufës së grimcave PSO

Inteligjenca e grimcave siguron nje kuader per ndertimin dhe implementimin e sistemeve te bere nga shume agjente qe jane te afte te sjellin zgjidhjen e problemeve komplekse. Perparesite e kesaj teknike jane te shumte :

1) Fuqia kolektive : deshtimi i komponenteve individual nuk pengon punen apo ecurine e metodes.
2) Thjeshtesia individuale : sjellja bashkepunuese bent e mundur per te reduktuar kompleksitetin e individeve .
3) Shkallezueshmeria : mekanizmat e kontrollit te perdorur nuk varen nga numrii agjenteve ne tufe.

PSO eshhte nje teknike popullore qe bazohet ne inteligjencen e tufes se grimcave e paraqitur fillimisht nga Kennedy dhe Eberhart ne vitin 1995.Idea themelore e kesaj teknike eshte modelimi matematik dhe simulimi i aktiviteteve te kerkimit te ushqimit ten je tufe zogjsh (grimcave).Ne hapesiren shumepermasore ,cdo grimce ne tufe eshte zhvendosur ne drejtim te pikes optimale duke shtuar nje shpejtesi me pozicionin e saj .Cdo grimce ne PSO fluturon ne hapesiren e kerkimit me nje shpejtesi te pershtatshme qe eshte modifikuar ne menyre dinamike ne baze te pervojes se vet te fluturimit dhe pervojes se fluturimit te grimcave te tjera .Tri komponente te cilat ndikojne ne shpejtesine e nje grimce jane : momenti i inercise

,moment njohes dhe ai social.Komponenti inercial simulon sjelljen inerciale te zogjve per te fluturuar ne drejtimin e meparshem.Komponenti njohes modelon kujtesen e zogut rreth pozicionit te tij te meparshem me te mire.Komponenti social modelon kujtesen e zogut rreth pozicionit me te mire ne mesin e grimcave .Pozicioni me i mire i marre deri tani nga cdo grimce eshte i njohur si *pbest* me i miri i pergjithshem nga te gjithe grimcat ne popullate njihet si *gbest.*

Cdo grimce perfaqeso nje zgjidhje te mundshme per nje problem me d-dimensione dhe genotipi i tij perbehet nga 2*d parametra.Se pari d-parametrat perfaqesojne pozicionet e grimcave dhe d-parametrat e tjere paraqesin komponentet e shpejtesise .Keto parametra levizin me nje shpejtesi te adaptueshme ne kuader te hapesires se kerkimit dhe mbajtjes se kujteses se vet ne poziten me te mire qe ka arritur ndonjehere.Parametrat ndryshojne kur leviz nga nje iteracion ne tjetrin.Ne cdo iteracion ,funksioni fitness ,si mase e cilesise eshte llogaritur duke perdorur vector funksionin e tij.Pozicionin i seciles grimce ndryshon drejt *pbest*, dhe *gbest* bazohet ne shpejtesine e grimces e cila merret ne iteracionin tjeter si me poshte:

$$V_i^{k+1} = wV_i^k + c_1 r_1\left(pbest_i^k - X_i^k\right) + c_2 r_2\left(gbest^k - X_i^k\right) \qquad 4.2$$

$$X_i^{k+1} = X_i^k + V_i^{k+1} \qquad 4.3$$

Ku *w* eshte pesha inerci dhe $c_1\ dhe\ c_2$ jane konstante nxitim qe udheheqin grimcat drejt pozitave te permiresuara ,$r_1\ dhe\ r_2$ jane variabla te rastesishme te shperndara ne menyre uniforme midis 0 dhe 1 ; *k* tregon evolucionin e iteracionit .Algoritmi PSO perfundon pas nje numri maksimal iteracionesh ose kur pozicioni me i mire i grimcave te tufes nuk mund te permiresohet pas nje numri te mjaftueshem iteracionesh. Sfondi historic me zhvillimin dhe perpunimin e perafrimit te tufes se grimcave eshte shqyrtuar me detaje nga Banks ne [236].Hulumtuesit kane propozuar gjithashtu shume variante te PSO sic jane PSO diskrete, PSO Gausiane ,hibritizimi i GA dhe PSO , PSO multiobjektiv , dhe vleresuesi vektorial PSO .Disa nga keto variante jane propozuar per te perfshire aftesite e teknikave llogaritese (si versionet hibrite te PSO),ose per pershtatjen e parametrave te PSO per nje performance me te mire.Ne raste te tjera ,natyra e problemit qe te zgjidhet kerkon qe PSO te punoje ne mjedise komplekse si ne rastin e problemeve multi-objektive ose problemet ekufizuara optimizimi.

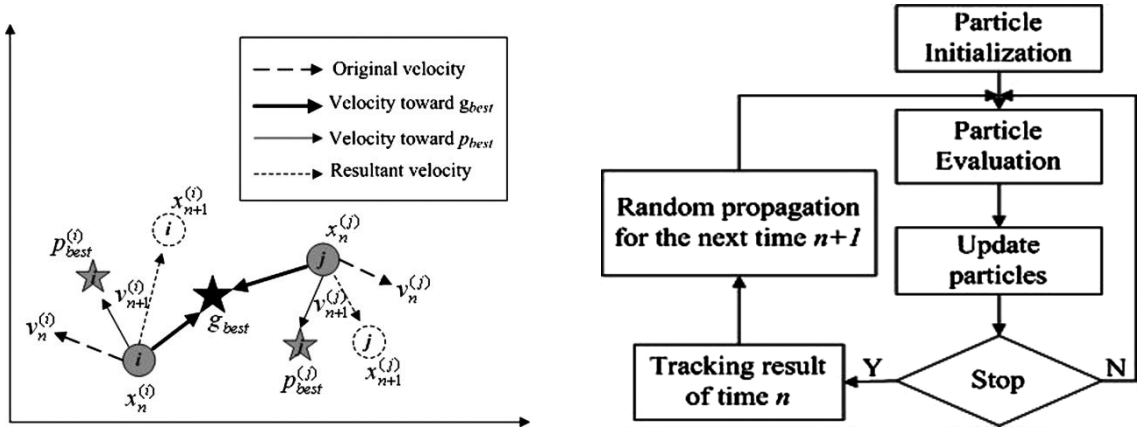


Figura 4.1 Puna për përmirësimin e algoritmit PSO

4.3 Formulimi i problemit

Do te formulohet problemi duke kryer optimizimin e bankes se filtrave me ane te teknikes PSO .Funksioni objektiv i propozuar qe do te perdoret per algoritmet evolucionare perbehet nga nje klasifikues foneme,dhe performance e ketij klasifikuesi eshte vlera fitness per vleresuesin individual.

Për njohjen e sinjaleve të ndryshme audio dhe për ti përdorur në mënyrën e duhur duhet të përshtatet metoda e njohjes së fjalës.Përqindja e gabimit është diferenca midis outputit actual dhe outputit të dëshiruar i llogaritur mbi bazën e teknikave të ndryshme.

Pas analizimit të sinjalit ,ai njihet nga sistemi në proçesin e testimit.Proçesi bazë në sistemin e njohjes është nxjerrja e karakteristikave apo tipareve,trajnimi dhe testimi i sinjalit të të folurit ,këtu trajnimi dhe testimi është bërë nga ANN (rrjetat neural artificiale).ANN është përdorur në fazën e trajnimit dhe të njohjes.

4.4 Rrjeta neurale artificial ANN

Rrjeta neural artificial janë modele të të mësuarit të cilat janë frymëzuar nga rrjeta neural biologjike(sistemi nervor i njeriut) ,dhe janë përdorur për të përafruar funksionin.ANN përbëhet nga shumë nyje të cilat janë quajtur dhe element të përpunimit .Nyjet shtrihen të lidhura me lidhje të peshuar.ANN mund të jenë me dy shtresa ose me shumë shtresa .ANN me dy shtresa ka njësi input dhe njësi output.ANN me dy shtresa ka njësinë input ,njësinë e

fshehur dhe njësinë output .Në fillim të dhënat futen në njësinë e fshehur më pas shkojnë në njësinë output për përpunimin final.Në ANN me shumë shtresa lidhja midis inputit dhe outputit është jodirekte ,ndërsa në atë me një shtresë kjo lidhje është direkte.ANN ka tri pjesë bazë ,arkitektura e ANN ,të mësuarit ANN dhe testimi i ANN.Në arkitekturën e ANN janë shumë tipe të ANN në dispozicion si : Propagandimi kthim i ANN, ANN periodike ,ANN me funksion bazë radial,Harta e vetorganizimit ANN.Në të mësuarit e ANN kemi mësim me mbikqyrje dhe mësim pambikqyrje.Në mësimin pambikqyrje output nuk njihet dhe në mësimin me mbikqyrje njihet outputi .Në testimin ANN krahasohet output i ANN me outputin e dëshiruar.Nëse output nuk është i njëjtë me atë të dëshiruarin atëhere trajnimi i ANN është i nevojshëm të vazhdohet më tej dhe anasjelltas.

ANN e thjeshtë përbëhet nga tri shtresa : shtresa input ,shtresa e fshehur dhe shtresa output .Shumë inpute mund të ushqehen tek inputi i dhënë.Shtresa input ka një lidhje indirekte me shtresën e fshehur,dhe shtresa e fshehur ka një lidhje indirekte me shtresën output.ANN mund të përbëhet nga shumë nyje ose element të përpunimit .Diagrama e ANN jepet më poshtë.

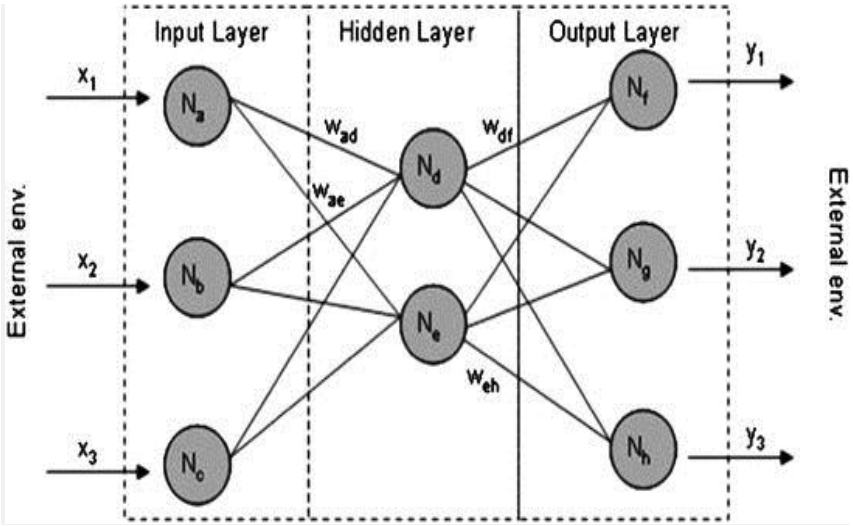


Figura 4.2 Rrjeta ANN

Për njohjen e fjalës janë përdorur metoda të ndryshme të ANN ,në mënyrë të ngjashme dhe për identifikimin e gjuhës janë përshtatur metoda të ANN .

Në këtë punim është përdorur njohja e fjalës dhe identifikimi i gjuhës me ANN , zgjedhja e tipareve apo karakteristikave me PSO dhe pa PSO .Pra objektivi kryesor këtu është ndarë në dy faza : Faza e pare është njohja e fjalës nga një sinjal i fjalës shqip dhe faza e dytë është

njohja e fjalës nga sinjali i fjalës shqip me kombinimin e përzgjedhjes tipareve PSO me ANN .Pasi aplikohen këto dy metoda krahasimi është bërë në bazë të normës së njohjes dhe gabimit .

Rrjeta Neurale ANN,(në këtë punim është përdorur rrjeta e hartave Kohonen me vet organizim), ka një arkitekturë mjaft të thjeshtë në krahasim me rrjetat e tjera neurale. Në rastin e një rrjete neurale ANN mungon shtresa e fshehur si rrjedhojë përdoruesit i mbetet të përcaktojë vetëm numrin e neuroneve në shtresën hyrëse dhe atë dalëse. Për t‟i dhënë një kuptim

rrjetës neurale, mund të imagjinohet problem i gjetjes së daljes së saktë në rrjetën neurale si një proces klasterizimi.Kjo do të thotë se gjatë procesit të trajnimit, si pozicionohen pikat në një hapësirë N-Dimensionale (ku N është numri i neuroneve hyrëse në sistem), bëhet një përafrim i këtyre pikave në klastera, në grupe të mëdha pikash të tilla që pikat brenda një klasteri janë më pranë njëra tjetrës se me çdo pikë tjetër të gjendur në një klaster të ndryshëm. Duke pasur parasysh se koeficientët MFCC si elementë hyrës në sistemin që po ndërtohet janë 12, mund të konsiderohet se secila prej këtyre 12 vlerave përcakton në mënyrë të pavarur një nga 12 parametrat përcaktuese të frekuencave karakteristike të sinjalit, pra i cakton një numër real secilit prej këtyre 12 komponentëve, atëherë hapësira në të cilën do të shpërndaheshin pikat gjatë trajnimit do të qe 12 dimensionale.

Gjithësesi problemi i vërtetë qëndron në identifikimin jo të neuroneve në hyrje në rastin e ndërtimit të ANN por të neuroneve në dalje.

Duke u nisur nga analizat e paraqitura për modelin HMM, shpesh herë për një karakter të vetëm përcaktohen tre gjendje të ndryshme. Një është gjendja hyrëse në fonemë (f1), e dyta është faza tranzitore ku korniza ka kapur momentin gjatë së cilës po zhvillohet ligjërimi i fonemës (f2) dhe së fundmi faza e tretë apo ajo e daljes nga fonema ku pregatitet aparati i të folur për shprehjen e karakterit pasardhës (f3). Si rrjedhojë e kësaj analize do të duhej të ndërtohej një shtresë në dalje për rrjetin ANN i cili të përfshinte për secilën nga 36 shkronjat e alfabetit shqip nga 3 gjendje, të cilat së fundmi do të konvergjonin në një fonemë të vetme. U përfshinë përkatësisht edhe 2 neurone që të mblidhnin momentet e heshtjes mes ndarjeve të neuroneve. Nga rezulatet eksperimantale u vu re se në modelin me 1 * 36 shkronja + 1 hapësirë në dalje, pra 37 neurone modeli konvergjon.

Si rrjedhojë e analizave të mësipërme u vendos që sistemi të jetë i ndërtuar duke përdorur nga njëra anë një shtresë hyrje të përbërë nga 12 neurone dhe një shtresë dalëse të përbërë nga 37 neurone në rastin e të gjithë shkronjave të gjuhë shqipe.Numrat përmbajnë vetëm 24

shkronja,atëherë për ndërtimin e modelit për numrat u përdor po e njëjta rrugë si për shkronjat, pra ajo e zgjedhjes së 24 nyjeve në dalje (në fakt numrat në gjuhën shqipe përmbajnë 23 shkronja + 1 dalje për pauzën atëherë 24 dalje). Si përfundim modeli konvergjoi, siç duket edhe nga imazhi më poshtë, në vlerën e 0.2775 të gabimit dhe kjo që në iteracionin e 4-t të trajnimit. Më pas kjo vlerë mbetet konstante.

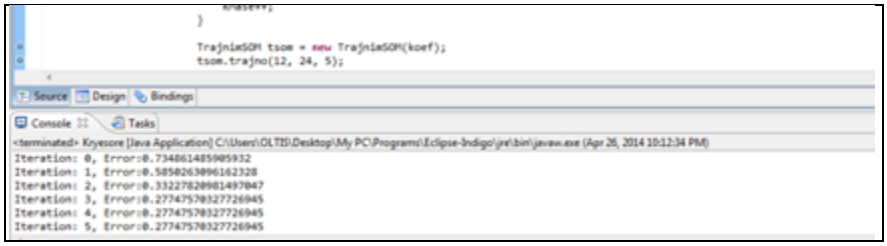


Figura 4.3 : Modeli konvergjent për njohjen

4.5 Trajnimi i rrjetës neurale

Procesi i trajnimit të rrjetës neurale është një nga proceset kyç. Në fakt vetë ky proces në rastin e trajnimeve të pasupervizuara dhe në veçanti në rastin e ANN-së ndahen në dy pjesë: i) trajnimi i peshave të rrjetit neural dhe ii) identifikimi i vlerave përkatëse për vlerat e paracaktuara në hyrje. Pra thënë më saktë; fillimisht trajnohet sistemi me mënyrën standarde dhe më pas provohen vlerat për të gjetur një motiv brenda rezultateve në hyrje pra duke krahasuar rezultatet e trajnimit me ato në dalje.

Vetë procesi i trajnimit kryehet në iteracione të cilat varen nga trajnimi në trajnim. Numri i këtyre iteracioneve mund të jetë i ndryshëm në sisteme të ndryshme apo në të njëjtin sistem për parametra të ndryshëm. Hap pas hapi llogaritet nga sistemi një parametër gabimi i cili duhet të vejë duke u zvogëluar vazhdimisht që sistemi të quhet i pranueshëm (pra sistemi duhet të konvergjojë dhe ta kryejë këtë pa oshilime rreth rrezultatit të ekuilibrit).
Gjatë prodhimit të koefiçientëve MFCC, numri i koefiçientëve të prodhuar për çdo njësi kohore është përgjithësisht 12, 13 ose 39. Në varësi të këtij numri edhe numri i neuroneve në shtresën hyrëse do të jetë 12, 13 apo 39 përkatësisht. Në rastin e këtij dizertacioni, për të ulur kompleksitetin e informacionit hyrës, numri i neuroneve në hyrje do të jetë 12.
Në shtresën dalëse sistemi duhet të ketë aq neurone sa është edhe numri i mundshëm i rezulateve të ndryshme që mund të marren nga kombinimet e informacioneve në neuronet

hyrëse. Duke qenë se gjuha shqipe ka 36 shkronja dhe po aq fonema atëherë po 36 duhet të jetë edhe numri i neuroneve në dalje që do të përdoreshin nga rrjeti neural për zgjidhjen e plotë të problemit të Njohjes së fjalës.

Algoritmi i përdorur për trajnimin e këtij sistemi është mjaft i thjeshtë dhe me pak hapa. Përkatësisht hapat mund të përmblidhen në: 1) Inicializimi, 2) Kampionimi, 3) Zgjedhja e neuronit fitues, 4) Përditësimi i ndryshimeve (Δ) dhe 5) Përsëritja e hapavenga 2 deri në 4 derisa Δ = 0. Këto janë edhe hapat bazë që ndjek çdo algoritëm trajnimi. Por ndryshe nga çdo model tjetër trajnimi, ANN ka një veçanti. Neuronet në rastin e ANN konkurojnë me njëra-tjetrën për të gjetur një fitues, që është pikërisht neuroni me distancë minimale euklidiane me nyjet në hyrje.

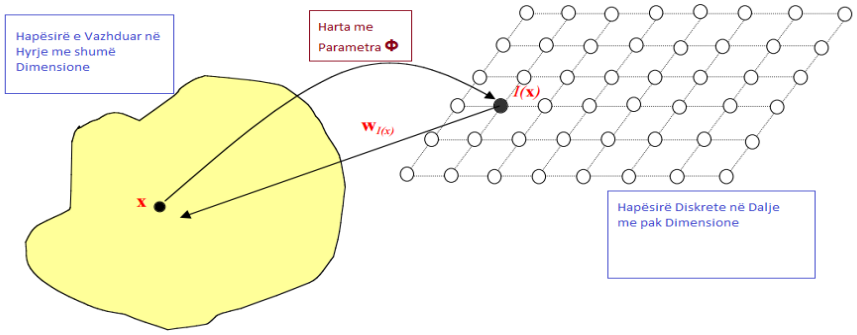


Figura 4.4. Skemë funksionale e rrjetës Neurale

Në vijim po shpjegojmë hapat për kryerjen e këtyre veprimeve duke përshkruar proceset që marrin pjesë në trajnimin e këtij sistemi. Për të kryer këtë, po supozojmë se kemi një rrjet neural pa shtresë të fshehur, me N nyje në hyrje dhe M nyje në dalje dhe ku rrjeti është i lidhur në mënyrë të plotë, pra çdo nyje në hyrje lidhet me të gjitha nyjet në dalje. Shpesh herë këtyre rrjeteve në literaturë ju referohen edhe si Rrjete Kohonen.

Le t"i numërojmë me i ϵ [1, N] neuronet në hyrje dhe me j ϵ [1, M] neuronet në dalje. Me w_{ij} shënojmë peshat që lidhin neuronin në hyrje N_i me atë në dalje N_j.

1. Inicializimi

Fillimisht inicializohen me vlera të rastësishme peshat w_{ij} që lidhin neuronet i dhe j. Në vijim këto pesha do të përmirësohen vazhdimisht për të arritur në rezultatet e dëshirueshme.

2. Procesi kompetitiv

Procesi kompetitiv është procesi i konkurimit të neuroneve dalëse për të marrë rezulatet hyrës. Mënyra si konkurojnë mes njëra tjetrës neuronet dalëse është duke llogaritur distancën Euklidiane mes nyjeve hyrëse dhe peshave w_{ij} që tregojnë lidhjen mes gjithë nyjeve hyrëse dhe një nyje dalëse.

$$d_j(x) = \sum_{i=1}^{D} (x_i - w_{ji})^2$$

Mes gjithë rezultateve të marra, ai që do të zgjidhej është pikërisht rezultati me distance euklidiane më të vogël.

3. Procesi kooperativ

Kjo lloj rrjeteANN ka një veçanti në procesin e trajnimit. Përveç neuronit fitues i cili do të ketë mundësinë të përmirësojë peshat e tij. Peshat nuk do t‟i ndryshojë vetëm neuroni qëndror, por edhe të gjithë neuronet që janë pranë tij sipas një afërsie topologjike me neuronin qëndror. Sa më pranë të jetë një neuron me atë që është fituesi, aq më i madh është ndryshimi. Ky funksion është simetrik dhejep një shtrihet gradualisht mbi nyjet.

$$T_{j,I(x)} = \exp(-S_{j,I(x)}^2/2\sigma^2)$$

$$\sigma(t) = \sigma_0 \exp(-t/\tau_\sigma)$$

4. Procesi adaptiv

Procesi adaptiv është një proces tepër i rëndësishëm i cili tregon masën e ndryshimit të peshave mes neuroneve të ndryshëm. Edhe ky funksion si ai I mëparshëm lidhet me një faktor i cili ndryshon me kalimin e kohës. Ky funksion jep ndryshime bazuar në një transformim kohor i cili ndryshon me kalimin e kohës. Për më tepër vlera e kalkuluar e Deltas të këtij funksioni do t‟u aplikohet vlerave të peshave. Në rast se kjo vlerë bëhet 0, algoritmi duhet të pushojë me iteracionet e tij.

$\Delta w_{ji} = \eta(t) \cdot T_{j,I(x)}(t) \cdot (x_i - w_{ji})$

$\eta(t) = \eta_0 \exp(-t/\tau_\eta)$

Ky algoritëm është mjaft efektiv dhe ngjan së tepërmi me mënyrën si funksionon truri i njeriut dhe mbi të gjitha me parimin e ndarjes së informacionit në tru në pjesë të ndryshme për të ndihmuar me procesin e njohjes dhe të mësuarit.

Metoda apo përafrimi i propozuar për trajnimin e rrjetës neurale që është përdorur në këtë punim ,është metoda që përdor përoptimizimin e peshave metodën PSO.Hapat që ndiqen për trajnimin e rrjetës duke përdorur PSO janë:

Hapi 1) Mbledhjen e të dhënave

Hapi 2) Krijimi i rrjetit

Hapi 3) Konfiguro rrjetin

Hapi 4) Inicializimi i peshave dhe animeve

Hapi 5) Trajnimi i rrjetit duke përdorur PSO

Hapi 6) Vleresimi i rrjetin

Hapi 7) Përdorimi i rrjetit

Rrjetat neurale paraqesin një arkitekturë të thjeshtë ndërtimore, të ngjashme me strukturën neurale të njeriut (të paktën kështu pretendohet se sillen) por të cilat shpesh herë duan një kapacitet të lartë llogaritës dhe algoritme të shpejtë. Jo rrallë herë, performanca e rrjetit neural varet në mënyrë të drejtpërdrejtë nga koha e kryerjes së llogaritjeve. Duke qenë se llogaritjet që kryhen për përcaktimin e peshave të lidhjeve mes neuroneve të një rrjeti janë njehsime matricore, një algoritëm jo performues i shumëzimit, mbledhjes, renditjes apo transpozimit të matricave etj. mund të afektonte në mënyrë të drejtpërdrejtë kohën e përgjigjes dhe vlefshmërinë e të dhënës.Prandaj në këtë punim është përdorur algoritmi PSO si dhe përmirësimi i tij i kombinuar me rrjetën neurale për të dhënë një rezultat të përmirësuar të njohjes së fjalës dhe në një kohë optimale.Nuka ka një funksion matematikor që lidh ID e tingujve dhe kohën të shprehur në sekonda ,për të ekzekutuar algoritmin PSO .Metodologjia konsiston në vënien si një pozicion variable të identitetit të tingullit kundrejt parametrit të kohës që i nevojitet algoritmit PSO për ekzekutim dhe algoritmit të përmirësuar apo zgjeruar PSO.Nëpërmjet Matlabit këto të dhëna analizohen permes analizës së regresit linear duke dhënë koefiçientët e përcaktueshmërisë.Në kapitullin 5 janë paraqitur rezultatet e analizës së regresit linear.

KAPITULLI V

Aplikime dhe Rezultate

5.1 Demostrimi i sjelljes së metodave të parashikimit linear gjatë aplikimit të tyre në njohjen e fjalës

Metodat e mesiperme te parashikimit linear jane studiuar dhe krahasuar nga pikepamja e zbutjes se zhurmes se shtuar dhe aplikimit ne njohjen e fjales me implementim ne Matlab.Ne vend te metodes se parashikimit WLP eshte perdorur metoda SWLP .Kjo metode ka pole te qendrueshme prandaj performance e saj eshte me e mire .Vleresimi yne tregon fuqi spektrale dhume te mire per te njejtin sinjal te fjales .Ne menyre te vecante per SWLP rezultati eshte shume me I mire krahasuar me metodat e tjera .Per te njejten SNR marrim EER me te mire ,rrjedhimisht saktesia e ketij zbatimi eshte shume me e larte .Metoda zbulon ne menyre me te mire tiparet e spektrit dhe nuk kemi nje kompleksitet kompjuterik .Ne menyre qe te sigurohet stabiliteti I metodes WLP ne duhet te bejme nje kerkese shtese per te qene te sigurt nese te gjithe polet e modelit jane te shperndara Brenda rrethit njesi .Sic mund te tregohet dhe nga figurat e meposhtme nese e drejtojme metodat e mesiperme mbi nje sinjal diskret ne domenin kohe ,merren rezultate te ndryshme .Dallimet mes tre ploteve tregojne nje permiresim te dukshem nga parashikimi linear tradicional deri tek parashikimi linear I qendrueshem .Vlera e funksionit jep amplitude e sinjalit te fjales tipar.Informacioni I marre nga LP nuk eshte aq I detajuar krahasuar me metodat e tjera .Ploti I LP eshte I qete keshtu qe nuk zbulon shume informacion .Kjo ndodh sepse nuk jepet nje tablloe dallimeve te funksionit midis dy vlerave te ndryshme te frekuencave.WLP na jep nje informacion me te mire ne lidhje me karakteristikat .Nga ploti mund te shohim se karakteristikat jane me te hollesishme sepse ploti eshte me kodrinor ,dhe sic dihet ndryshojne ne varesi te frekuencave .

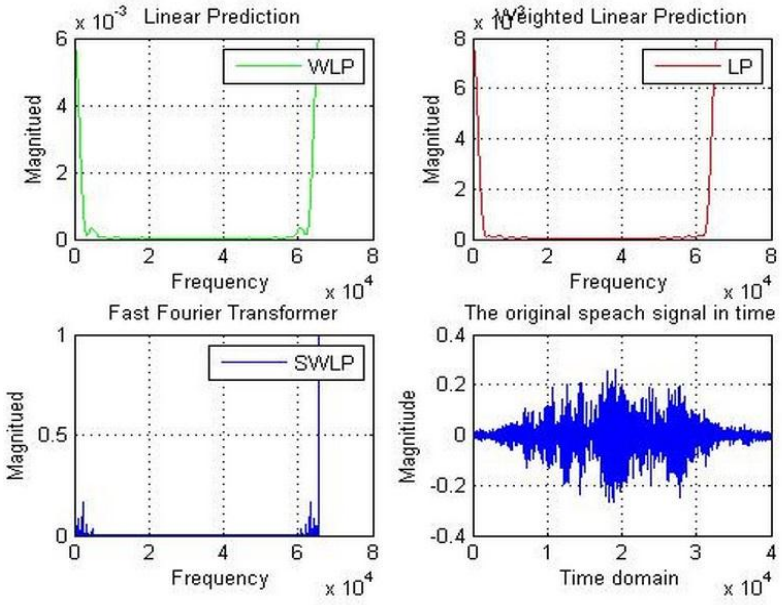


Figura 5.1 Spektrat FFT ,LP ,WLP për një zë të fjalës së shprehur

Meqe ne shume aplikime reale ,kemi prezencen e zhurmes atehere eshte e nevojshme te shihet efekti I zhurmes .Ne figuren e meposhtme paraqitet simulimi I sinjaleve te ndryshme te fjales me zhurma te ndryshme shtese.

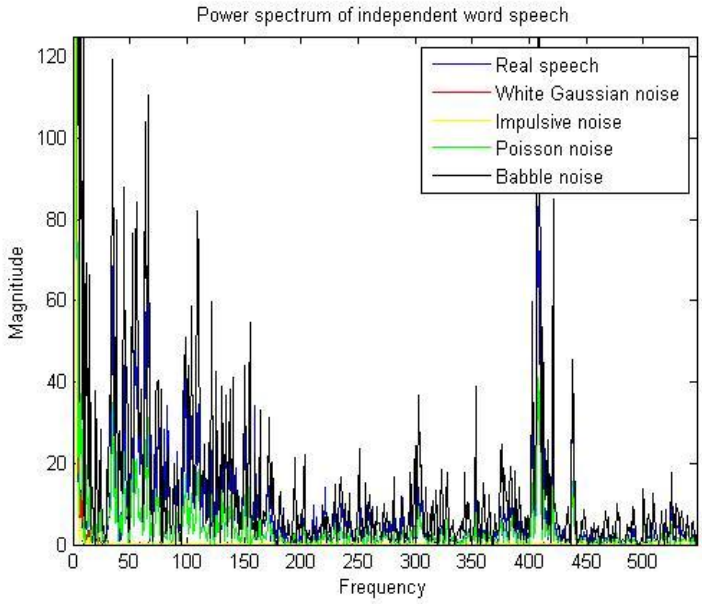


Figura 5.2 Fuqia e spektrit kundrejt zhurmave

Rezultatet e aritura me larte kane ardhe nga nje bashkesi te dhenash fillestare,I cili eshte nje sinjal arbitrar I regjistruar nga nje mikrofon I kompjuterit PC.Ne menyre q eta perdorim ate sin je hap perpunim duhet ta presim ate ne gjatesi te ndryshme ,te llogarisim spektrin nepermjet LP dhe te marrim mesataren e ketyre spektrave .Sinjale te tjera input qe duhet te testohen ne fund shfaqen ne nivele te ndryshme te zhurmes .Eshte shtuar zhurma Gausiane per te pare ndikimin e zhurmes .Lartesia e nivelit te zhurmes eshte e ndryshme ne spectra te ndryshem dhe kjo vjen nga pranimi i rreme apo aprovimi i rreme ,sic e tregon dhe algoritmi i ndertuar per njohjen e fjales duke kombinuar metodat e parashikimit linear .

Result: False or True

INITIALIZATION;

Average Speech Signal;

Speech Signal to be tested;

COMPUTE THE SPECTRUM;

Average Speech Signal Linear Prediction(LP);

Coeff=Training coefficient for the classifier;

while *The length of the speech signal spectrum* ***do***

Normalise the speech signal:

$Signal_{Normalized}=\frac{SpeechWeightedLP}{SignalMaxValueWeightedLP}$;

$Difference=Signal_{Normalize}WeightedLP - AverageSpeechSignalFFT;$

$Difference_{PCA}=PCA(Difference)_{first(s);}$

Dominant Component, the last one is noise;

Index=Classifier($Difference_{PCA}$, *Coeff*) if Index¡threshold **then**

Returne true;

else

Return false ;

end

end

Me poshte japim dhe performance e gabimit kundrejt metodave :

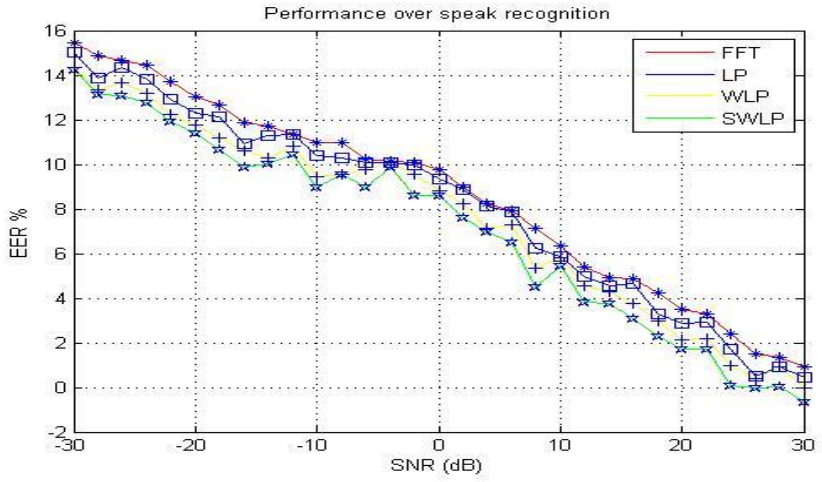


Figura 5.3 Performanca e gabimit kundrejt metodave

Norma e barabartë e gabimit (**EER**)							
MFCCs	Vlerësuesi Spektral	Norma e sinjali të zhurmës (dB)(**SNR**)					
		20	10	0	-10	-20	-30
Zhurma Gausiane	FFT	3.5	6.2	9.8	11	13	15.3
	LP	3 6	9.2	10.2	12.3	15	
	WLP	2.2	5.9	8.3	8.9	12	14.1
	SWLP	2	5.8	8.2	8.5	10.8	14

Tabela 5.1 : Rezultatet e prishjes së fjalës kundrejt zhurmës

Tabela tregon rezultatet e zbulimit të fjalës për zhurmën Gausiane,dmth normën e gabimit të prishjes së fjalimit nën ndikimin e zhurmës Gausiane.Rezultatet janë paraqitur duke përftuar karakteristikat MFCC nga përdorimi i teknikave LP,WLP, dhe SWLP. Nga tabela vërejmë që në nivelet e zhurmës -20,-30 ku performance e sistemit degradon me shpejtësi ,karakteristikat më rezistente janë MFCCs të përftuara nga WLP dhe SWLP .

Sistemi i propozuar në këtë punim për njohjen e fjalës bazohet në teknikën e njohjes audio të nxjerrjes së tipareve MFCC .Kjo teknikë u plotësua me disa teknika (LP,WLP,SWLP) për të

përmirësuar fuqinë e sinjalit në kushte të pafavorshme të zhurmës .Në veçanti vlerësuesi tradicional i spektrit Furie në llogaritjen e MFCC është zëvëndësuar me metoda të reja që kombinojnë modelimine mbështjellëses spektrale të parashikimit linear me një structure të mire spektrale ,dmth frekuencën themelore dhe harminikat e saj.

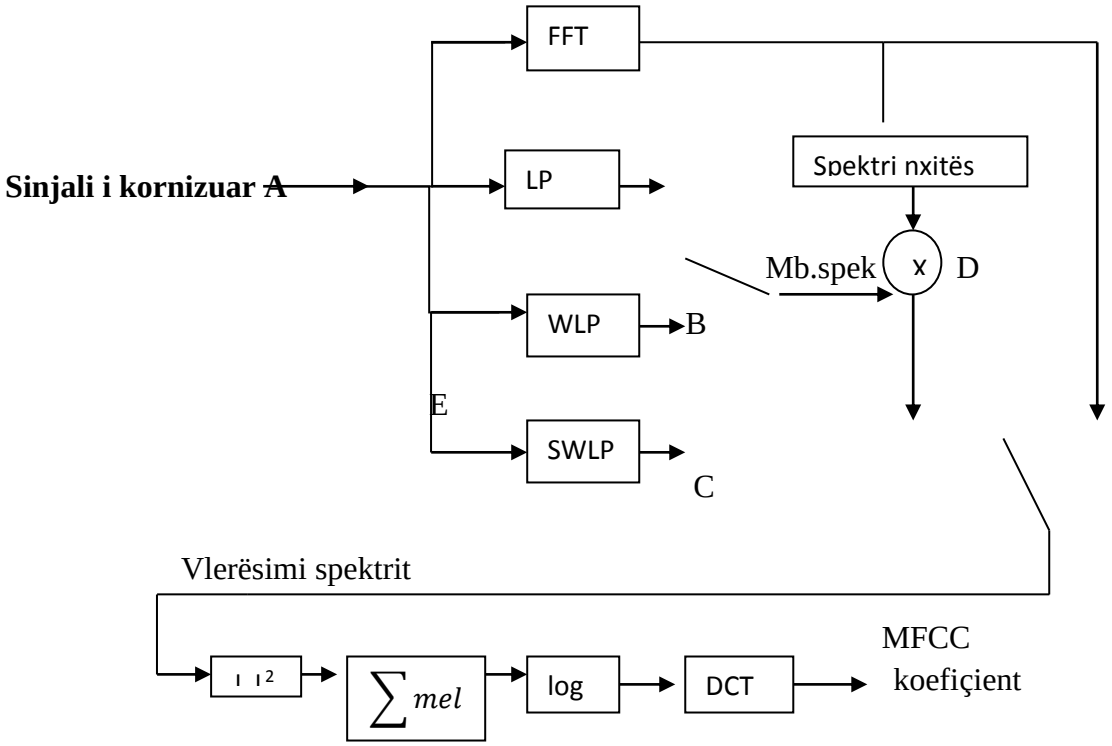


Figura 5.4 : Diagrami i propozuar i nxjerrjes së tipareve MFCC

FFT mbron mbeshtjellësen spektrale ,ndërkohë LP,WLP,SWLP përshkruajnë mbështjellësen spektrale .Për të përfshirë mbështjellësen spektrale ,mbështjellësja spektrale all-pol shumëzohet me spektrin ngacmim të marrë nga rritja e peshës së cepstrumit bazuar tek FFT në domenin cepstral dhe duke transformuar cepstrumin e peshuar(rritur në peshe) përsëri në domenin spectral.Mekanizmi lëvizës i përdorur në këtë rast vlerëson me zero koefiçientët cepstral që i korespondojnë intervaleve të kohës më të vegjël se (F_s/700)+1 ,ku F_s tregon shkallën e mostrave në Hz.Kjo do të thotë se informacioni i nxitjes periodike të traktit vocal deri në 700 Hz mbahet apo ruhet në spektrin spektrin nxitës .Pasi një vlerësimi spektri është marrë duke përdorur disa nga metodat e mësipërme, ai ngrihet në katror dhe ka kaluar nëpërmjet një banke konvencionale mel filtër me 40 filtra Mel. Pas konvertimin të

spektritmel në një formë logaritmike ,ai kalon nëpërmjet transformimit kosinus diskret ,DCT ,për të arritur në numrin e dëshiruar të korfiçientëve MFCC.

5.2 . Përmirësimi i njohjes së fjalës duke përdorur Rrjetat Nervore Neurale dhe teknikën e optimizimit PSO

PSO është një teknikë popullore që bazohet në inteligjencën e tufës së grimcave e paraqitur fillimisht nga Kennedy dhe Eberhart në vitin 1995.Idea themelore e kësaj teknike është modelimi matematik dhe simulimi i aktiviteteve të kërkimit të ushqimit të një tufe zogjsh (grimcave).Në hapësirën shumëpërmasore ,cdo grimcë në tufë është zhvendosur në drejtim të pikës optimale duke shtuar një shpejtësi me pozicionin e saj .Cdo grimcë në PSO fluturon në hapësirën e kërkimit me një shpejtësi të përshtatshme që është modifikuar në mënyrë dinamike në bazë tëpërvojës së vet të fluturimit dhe përvojës së fluturimit të grimcave tëtjera .Tri komponentë të cilat ndikojnënë shpejtësinë e një grimce janë : momenti i inercisë ,moment njohës dhe ai social.

Në shkenca kompjuterike, optimizimi i tufës së grimcave (PSO) është një metodë kompjuterike që optimizon një problem në mënyrë iterative duke u përpjekur për të përmirësuar një kandidat zgjidhje në lidhje me një masë të caktuar të cilësisë. PSO optimizon një problem duke pasur një popullsi prej zgjidhjeve kandidate, grimcat e koduara këtu, dhe duke lëvizur këto grimca rreth në hapësire-kërkuese në bazë të formulave të thjeshta matematikore mbi pozicionin e thërrmijës dhe shpejtësisë. Lëvizja e çdo grimce është ndikuar nga pozicioni i saj lokal më të mirë të njohur, por, është udhëzuar edhe drejt pozicioneve më të njohura në hapësirën-kërkuese, të cilat janë të përditësuar si pozicionet më të mire që janë gjetur nga grimcat e tjera. Kjo pritet të lëvizë tufën drejt zgjidhjes më të mira.

Do të përshkruajmë disa detaje të veçanta për aplikimin në rastin e përdorimit të teknikave të rrjetës neurale dhe teknikës optimizimit PSO ,për krijimin e një skedari të njohjes së fjalës duke identifikuar me një numër konkret .Raporti teknik jepet në disa detaje duke përdorur gjuhën e programimit Matlab .

Rezultatet

Fillimisht duhet të ekzekutohet menuja për njohjen e fjalës **(speechrecognition.p**) .

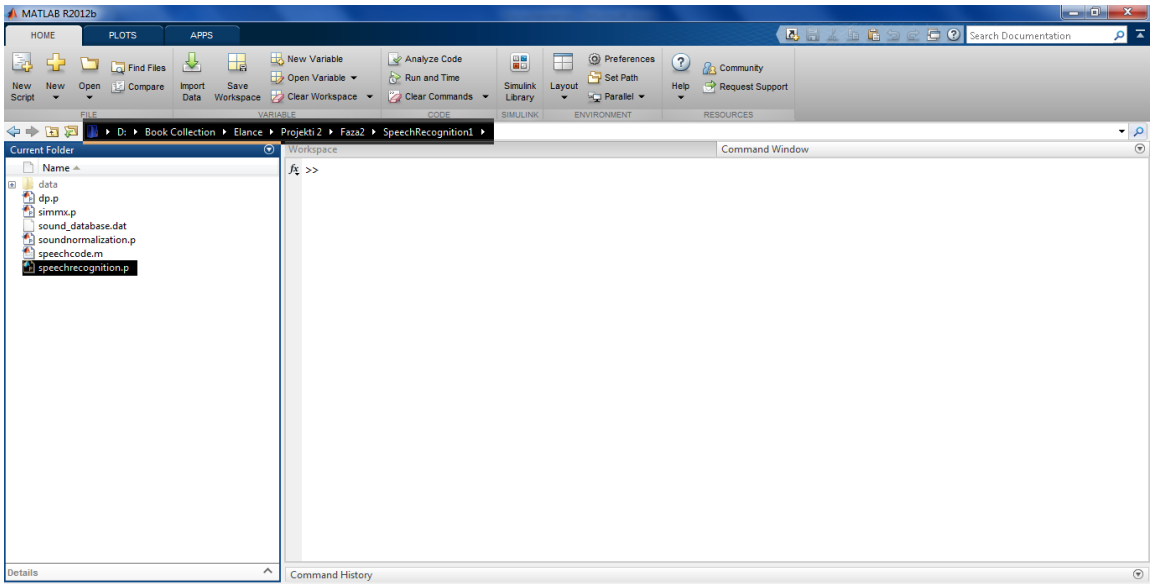


Figura 5.5 : Zgjedhja e file speechrecognition.p

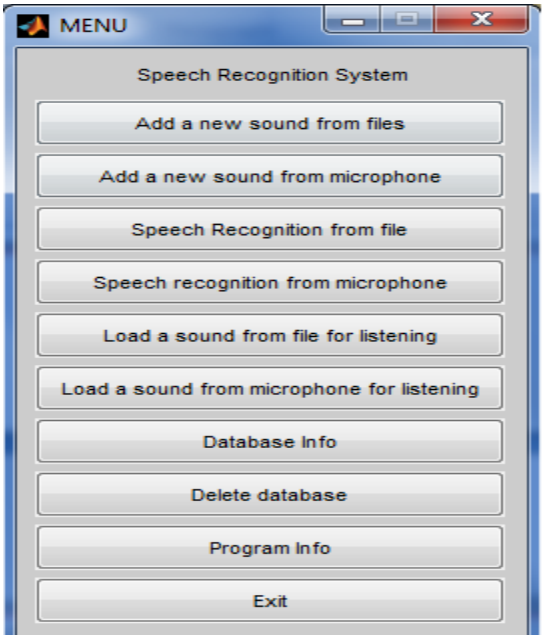


Figura 5.6 : Ekzekutimi i file speechrecognition.p

Dosja ka 10 dritare të vogla dhe secili prej tyre ka një funksion të veçantë ,le ti përshkruajmë si më poshtë :

Add a new sound from files : Merr dosjen e zërit nga kompjuteri dhe e vendos atë në databazë

Add a new sound from microphone : Regjistron zërin nga mikrofoni dhe e vendos atë në databazë .

Speech recognition from file : Kryen identifikimin e tingullit të marra nga kompjuteri.

Speech recognition from microphone : Kryen funksionin e identifikimit të tingullit të marrë nga mikrofoni i kompjuterit.

Load a sound from file for listening : Aktivizon(luan) dosjen e marrë nga kompjuteri .

*Load a sound from microphone for listening:*Aktivizon (luan) dosjen e marrë nga mikrofoni I kompjuterit.

Database info : Jep informacion mbi elementët faktik të përfshirë në databazë .

Delete database : Boshatis databazen .

Program info : Jep informacion për programin .

Exit : Mbyll programin

Zgjedhja e një file nga folder i të dhënave “data” ekzekutohet si më poshtë :

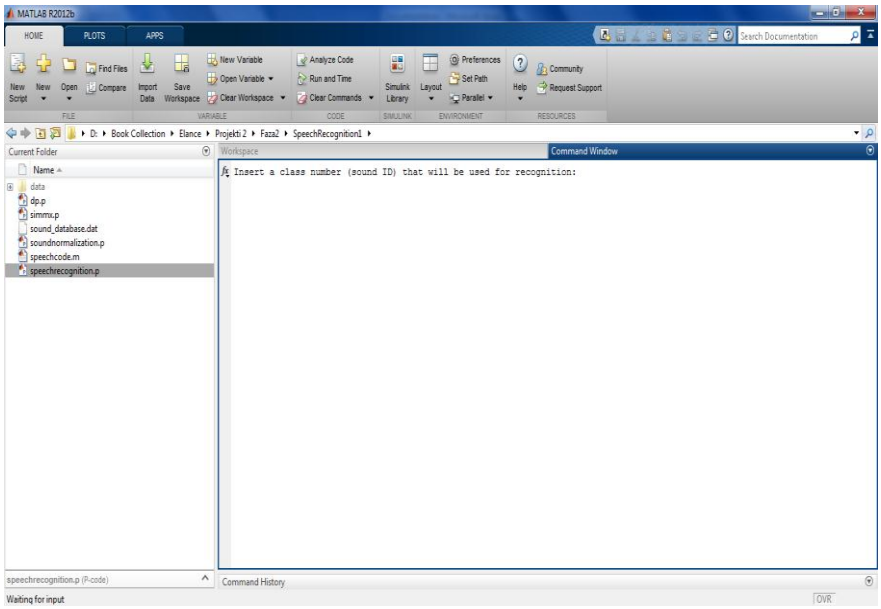
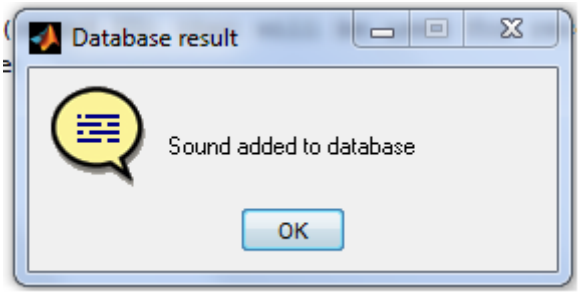


Figura 5.7: Shfaqja e dritares kur zgjidhet një file nga folder “ data “

Siç shihet nga komanda e shpejtë ,është e nevojshme që të vendoset një ID për dosjen në fjalë .Për thjeshtësi vendosim ID=1 . Pra ,në komandën e shpejtë shtypim numrin 1.Pasi futet numri ID=1 shfaqet një dritare si më poshtë:



Në mënyrë të ngjashme veprojmë me të gjithë file e tjerë që janë të përfshira në databazë.Për të pare se çfarë përmban databaza ,meqë ngarkojmë dosjet e tingujve , shtypim *Database info.*

Përshkrimi i dosjeve të aplikimit

Më poshtë po japim përshkrimin e dosjeve të aplikimit të cilat përbëhen nga dosjet burim.

1.Speechrecognition.m – Ky është funksioni kryesor i cili bën të mundur ekzekutimin e funksioneve të tjera .

2. Dabu.m - Ky është funksioni i cili bën të mundur regjistrimin e zërit dhe është i aktivizuar kur ne duam që të hyjmë në databazën e marrë nga mikrofoni .

3. Untitled.m – Ky është funksioni i cili përbëhet nga algoritmi PSO i përmirësuar .Kodi i të cilit konsiston në përmirësimin e PSO dhe paraqitet si më poshtë :

```
for iter = 1:maxIter
    swarm(:, 1, 1) = swarm(:, 1, 1) + swarm(:, 2, 1)/1.3;       %përditësimi i pozicionit x
me shpejtësi
    swarm(:, 1, 2) = swarm(:, 1, 2) + swarm(:, 2, 2)/1.3;       %përditësimi i pozicionit y
me shpejtësi

    x = swarm(:, 1, 1);                                         % marrja e pozicionit të
përditësuar
    y = swarm(:, 1, 2);                                         % pozicioni përditësuar
    fval = objfcn([x y]);                                       % vlerësimi funksionit duke
përdorur pozicionine grimcave

    % krahasimi vlerave të funksionit për të gjetur më të mirin
    for ii = 1:swarm_size
        if fval(ii,1) < swarm(ii,4,1)
            swarm(ii, 3, 1) = swarm(ii, 1, 1);                  % përditësimi pozicionit më
të mire x,
            swarm(ii, 3, 2) = swarm(ii, 1, 2);                  % përditësimi pozicionit më
të mire y,

            swarm(ii, 4, 1) = fval(ii,1);                       % përditësimi vlerës më të
mire deri tani
        end
    end
```

Algoritmi PSO :

Hapi-1: Nisja e popullsisë - pozicioni dhe shpejtësitë

Hapi-2: Vlerësimi i përshtatshmërisë së grimcës individuale(pbest)

Hapi-3: Ndiekja e individit me përshtatshmëri më të lartë (gbest)

Hap-4: Modifiko shpejtësitë bazuar në pBest dhe gbest pozicion.

Hapi-5: Update pozicionin e grimcave

Hapi-6: Përfundo në qoftë se kushti është plotësuar

Hapi-7: Shko tek Hapi 2.

Për të rritur performancën e rrjetës nervore neurale ANN ,është përdorur kjo teknikë optimizuese PSO ,duke optimizuar peshat ANN .Gjetja e një funksioni të rëndësishëm të përshtatshmërisë ose duke përdorur një lloj trajnimi diskriminues mund të përmirësojë rezultatet për të arritur performancë superior.Në PSO çdo grimcë përpiqet që të ndyshojë pozicionin e saj duke përdorur të dhënat e mëposhtme :- pozicionet e tanishme

-shpejtësitë aktuale

-distanca mes pozicionit actual dhe pbest

- distance mes pozicionit actual dhe gbest

Pozicioni actual (pika në kërkim në hapësirën e zgjidhjeve) mund të modifikohet nga ekuacioni i mëposhtëm : $S_i^{k+1} = s_i^k + V_i^{k+1}$ ku : V_i^{k+1} është shpejtësia e agjentit i në iteracionin e k+1 ,s_i^k është pozicioni actual i agjentit i në iteracionin e k .Në ekuacionin (4.2) është përdorur funksioni i mëposhtëm peshë:

w=wMax-[(wMax-wMin)iter]/maxIter ku : wMax =pesha fillestare

wMin=pesha përfundimtare

maxIter=numri max iteracioneve

iter= numri actual iteracioneve

Ndërtimi eksperimental

Në simulimin eksperimental janë përdorur dy database : një database trajnimi dhe një database testimi .Eksperimenti është testuar duke përdorur paketën Matlab.

Sinjali i zërit fillimisht është përpunuar duke hequr zonat heshtje,të cilat hiqen me anë të llogaritjes së energjisë me kohë të shkurtër .Janë zgjedhur gjithashtu segmente me 15ms.Një segment energjia e të cilit është më pak se rreth kufirit (pragut) krahasuar me energjinë mesatare të të gjithë sinjalit ,është hequr apo fshirë.Është aplikuar një filtër theksim i trajtës $H(z)=(1-0.95z^{-1})$ në sinjalin e fjalës.Sinjali i fjalës është ndarë pastaj në korniza analizuese ,ku sinjali është supozuar stacionar .Për vlerësim më të saktë të fjalës së shprehur ,frekuenca e marrjes së mostrave duhet të jetë më shumë se 6kHz për të përfshirë të paktën nje gjërësi fjalim prej 3kHz.Platforma themelore është një kompjuter PC i cili merr sinjalin e fjalës ,bën transformimin FFT për të nxjerrë karakteristikat ,"ushqen" me këto karakteristika një klasifikues të trajnuar dhe e kyhen identitetin e folësit në një mënyrë në kohë reale.Bashkësia e karakteristikave e përdorur për kodimin e fjalës ka një formë të koefiçientëve cepstral të llogaritur si më poshtë :

1.Ndarja e sinjalit të fjalës në veçimin e gjysmës së kornizave të 128 pikave.Kornizat gjysëem fqinje janë bashkuar për të formuar një kornizë të plotë të 256 pikëve.Meqë frekuenca e modulimit është 8kHz çdo kornizë zgjat 256/8=30msec.

2.Cdo kornizë është dritarizuar duke përdorur dritaren Hamming të 15 msec.

3.Llogariten koefiçientët cepstral për çdo kornizë duke përdorur formulën : $\text{cepstrum}\{\text{frame}\}=\text{FFT}^{-1}(\log|\text{FFT}\{\text{frame}\}|)$.

4. 12 koefiçientët e parë të cepstrumit janë marrë si vektor i tipareve të kornizës .

Vektori i tipareve fillimisht është trajnuar me ANN pastaj optimizohet duke përdorur PSO për njohjen e fjalës.

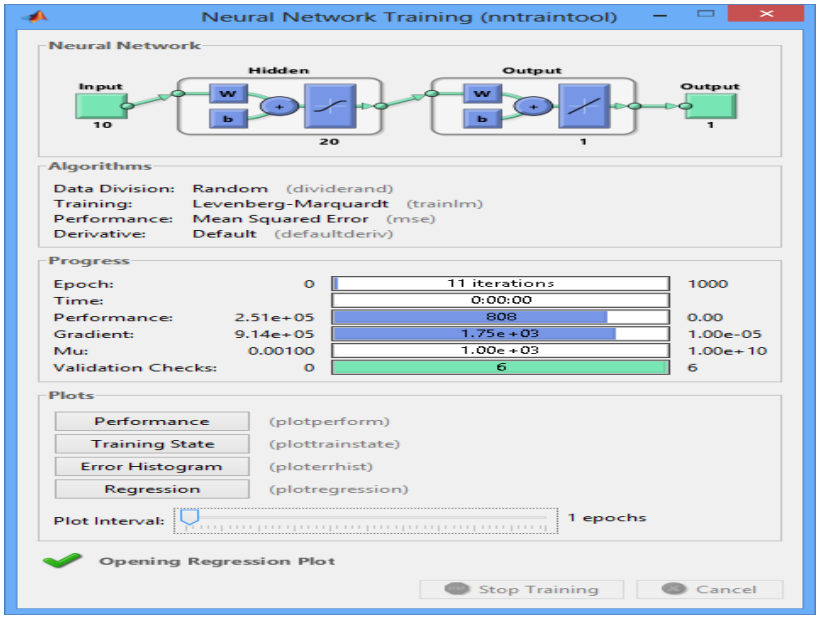


Figura 5.8 : Trajnimi i rrjetës neural

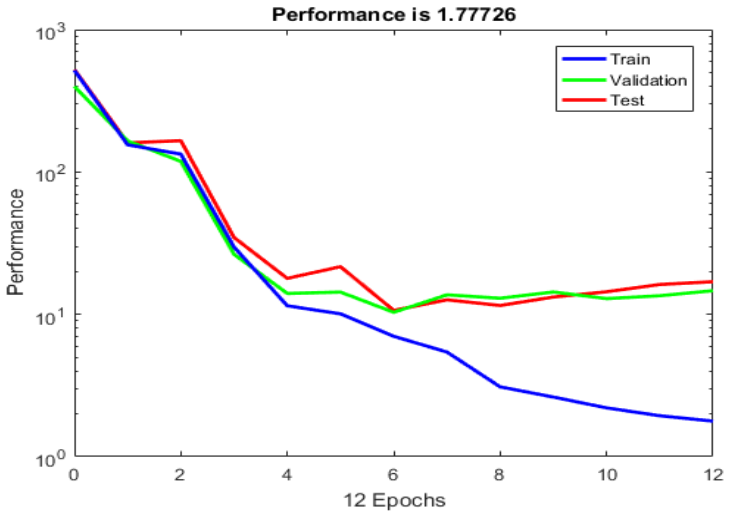


Figura 5.9 : Trajnimi i rrjetës neural

Pas trajnimit të rrjetës ,mund të kontrollohet performance e rrjetës që është përdorur për trajnimin dhe testimin e sinjaleve të fjalës të regjistruara në databazë..Figura e mësipërme tregon iteracionet apo periudhat kur vlefshmëria e performancës arrin një minimum.Siç shihet kurbat e vlefshmërise dhe testimit janë shumë të ngjashme ,që do të thotë njohja e sinjaleve të fjalës pothuajse është arritur ,në të kundër do të thonim që mund të ketë ndodhur ndonjë jopërshtatje.

Hapi tjetër për vlerësimin e rrjetës është krijimi i një ploti regresi i cili tregon marrëdhënien midis rezultateve (outputit) të rrjetës dhe objektivave .Nëse trajnimi i sinjalit të fjalës është perfekt ,rezultatet e rrjetës dhe objektivat do të jenë saktësisht të barabartë,por marredhenia rrallë është perfekte në praktikë.

Një kompleksitet që u shfaq gjatë trajnimit të sistemit ishte dhe dhe koha që i duhej këtij procesi për të përfunduar dhe kjo mund të vijë si rezultat i numrit të gabuar të nënshtrasave në shtresën e fshehur , si dhe për të shmangur problemin e zones së minimumit local ,prandaj për të shmangur këto u përdor algoritmi PSO sic u shpjegua më lart në fazën e trajnimit të rrjetës duke sjellë një përmirësim të njohjes së fjalës duke përdorur databazën e krijuar.Rezultatet e treguara dhe më poshtë tregojnë që rrjeta neural e trajnuar me PSO e përmirësuar ka nevojë për më pak iteracione dhe arrin saktësinë më të mire të trajnimit.

Puna e sygjeruar këtu njeh nëse sinjali fjalim është i njohur apo jo i njohur ,dmth nëse ID e zërit njihet ose jo .Saktësia e njohjes së të folurit është matur në aspektin e numrit të sinjaleve të të folurit të njohura ,kundrejt numrit total të sinjaleve të aplikuara për njohjen e sinjalit.

$$RR=\frac{numri\ i\ sinjaleve\ të\ njohura}{numrin\ total\ te\ sinjaleve\ të\ aplikuara} * 100$$

Gjithashtu performance mund të matet me perqindjen e gabimit : Nëse output actual është i ndryshëm nga rezultati i dëshiruar atëhere gabimi ndodh.

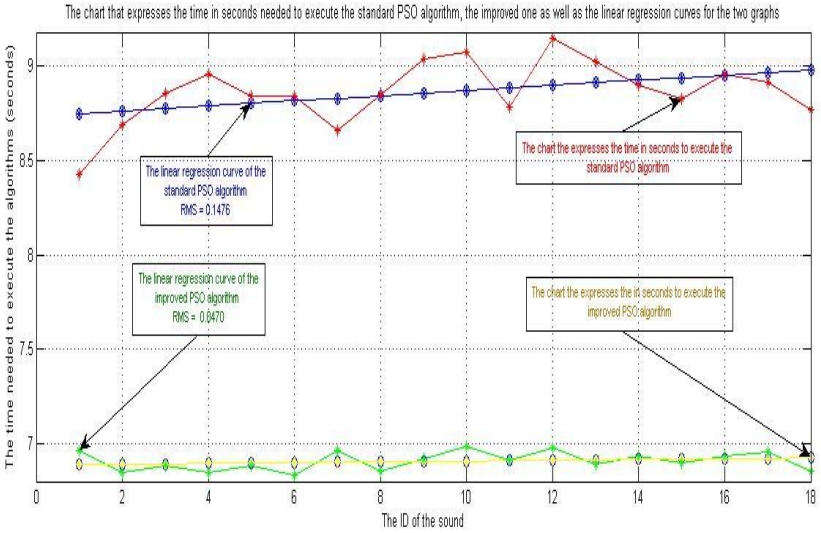


Fig 5.10 Ploti i regresit për trajnimin e ANN me PSO

Në fig 5.10 në pjesën e sipërme të figurës, koha e nevojshme për të ekzekutuar standardi PSO algorithm ka koeficientin e përcaktimit të barabartë me 0.1476, që do të thotë prej 14.76% të varësisë.

Në fig 5.10 në fund të figurës, koha e nevojshme për të ekzekutuar përmirësuar algorithm PSO ka koeficientin e përcaktimit të 0.047, që do të thotë nga 4.7% varësisë.

Theksohet që përmirësimi që është bërë konsiston në faktin se koha për të dhënë rezultat është rreth 30% më e ulët se algorithm standarde PSO.

Konkluzione

Për të kryer këtë punim doktorature raporti u nda në 5 kapituj :

- I pari është një prezantim i koncepteve teorike në prodhimin e të folurit ,përceptimin dhe analizimin .Kështu pjesa teorike pretendon të japë një sfond të rëndësishëm në lidhje me analizën e të folurit ,për të kuptuar parimet themelore ku është bazuar implementimi.
- I dyti prezanton studimin e metodave për vlerësimin e zhvillimit spectral të njohjes së fjalës ,të cilat janë metoda të përmirësuara të metodës së parashikimit linear .Është ndërtuar një algoritëm i njohjes së fjalës,konkretisht për vlerësimin e performancës në njohjen e zhurmshme të fjalës .Metodat e përmirësuara të parashikimit linear janë përdorur në fazën e njohjes së karakteristikave apo tipareve në rastin e një fjalimi apo zëri të regjistruar në kompjuter .Duke përdorur këtë janë krahasuar metodat e përmirësuar të parashikimit linear me metodën tradicionale FFT.Janë diskutuar metodat all-pol të parashikimit linear LP me fokus kryesor në peshën kohore të katrorit të mbetjes.Rezultatet e marra provojnë se pesha kohore në WLP ,e zbatuar duke përdorur funksionin STE ,çon në modele të spektrit që janë më pak të prirur nga zhurmat se sa spektrat të dhënë nga metodat konvencionale FFT dhe LP.
- I treti analizon gjuhën shqipe ,karakteristikat e saj .Ndërtohet një databazë e vogël e krijuar nga një lajm shqip i marrë në internet . Në aspektin e njohjes së fjalës shqipe ndërtohet një model i një rrjete neurale që në ndryshim nga meodeli HMM ,rrjetat nervore nuk bëjnë supozzime në lidhje me vetitë statistikore të karakteristikave dhe ka disa cilësi duke e bërë një model tërheqës.Duke pare që rrjetat neural kanë dhe mangësite e tyre si, mandësi në lidhje me kohën që i duhet për të trajnuar dhe testuar sinjalin e fjalës, pra shpejtësinë e trajnimit,për të shmangur këtë është përdorur teknika e optimizimit PSO .

Mendoj se duke qenë se çdo problem i përpunimit të gjuhës natyrore fillon meshqyrtimin e një fjalori dixhital, është detyrë e institucioneve shkencore të gjuhës shqipeta përpilojnë një fjalor të tillë dhe ta bëjnë të aksesueshëm për publikun pa pagesë. Njëpunë e tillë është bërë tashmë për të gjitha gjuhët e rëndësishme të botës dhe është njëdomosdoshmëri për gjuhën shqipe. Një fjalor i tillë i çertifikuar e lehtëson punën ekërkuesve dhe nuk lë rast për ekuivoke. Për më tepër, produkte të mbështetura mbi njëfjalor të pastruar mund të përdoren edhe si produkte komerciale.

Në përfundim të gjithë disertacionit mund të konkludojmë se avancimi dhezgjidhja e problemave të Përpunimit të të folurit Natyror është një hap i madh drejtautomatizimit të punëve njerëzore. Duke qenë se njeriu shkëmben me mjedisin që errethon rreth 25% të informacionit mes të folurit dhe të dëgjuarit, mjet bazë i së cilësështë gjuha natyrore, informatizimi i këtij procesi e afron më tepër njeriun me makinëndhe e bën këtë të fundit më familjare e të ngrohtë për përdoruesin e saj.

Referenca

[1] P. Strobach, Linear Prediction Theory-A Mathematical Basis for Adaptive Systems, Springer-Verlag, 1990.

[2] S. Haykin, Communication Systems, 4th Ed., John Wiley & Sons,2001

[3] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," Dig. Sig. Proc., vol. 10, no. 1, pp. 19–41, Jan. 2000.

[4] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P.Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," IEEE Sig. Proc. Lett., vol. 17, no. 6, pp. 599– 602, June [5] Handbook of speech recognition Benesty, Jacob; Sondhi, M. M.; Huang, Yiteng (Eds.) 2008.

[6]Pattern Recognition and Machine Learning C. Bishop,1 Feb 2007, ISBN-10: 0387310738, 2-nd edition [7]Makhoul,J ,Linear prediction a tutorial review, Procceding of the IEEE.

[8] Pattern Classification Richard O. Duda, Peter E. Hart, David G, November 9, 2000 | ISBN-10: 0471056693,2-nd edition .

[9] Understanding Digital Signal Processing (2nd Edition) by Richard G. Lyons.

[10]Fundamentals of Statistical Signal Processing, Volume III: Practical Algorithm, Development April 5, 2013 | ISBN-10: 013280803X .

[11] Pohjalainen ,J,Kallasjoki ,H,Plomaki ,K,J,Kurimo ,M and Alku, P., Weighted Linear Prediction for speech Analysis in Noisy Condition, in Proc .Interspeech ,Brighton ,UK ,2009.

[12]Saedi ,R,Pohjalainen ,J,Kinnunen ,Tand Alku ,P , Temporally Weighted Linear Prediction Features for Tackling Additive Noise in Speaker Verification , IEEE Signal Processing Letters 17(6)2010.

[13] Cucu,H., „Towards a speaker-independent, large-vocabulary continuous speech recognition system for Romanian" 44.

[14] Huang, X., Acero, A., Wuen-Hon., Chapter 11 „Language Modeling" in Spoken Language Processing- A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001 .

[15] [Juang], Juang, B.H., Rabiner, Lawrence R., „Automatic Speech Recognition – A brief History of the Technology Development", Georgia Institute of Technogy, Atlanta.

[16] Professor Dan Jurafsky, Lecture 4.6 Interpolation, Stanford NLP, https://www.youtube.com/watch?v=-aMYz1tMfPg, accessed at 01/06/2014 .

[17] Professor Dan Jurafsky, Lecture 4.3 Evaluation and Perplexity, Stanford NLP, https://class.coursera.org/nlp/lecture/129, accessed at 17/06/2014.

[18] Wikpedia, Perplexity, http://en.wikipedia.org/wiki/Perplexity, accesed at 29/05/2014.

[19] Huang, X., Acero, A., Wuen-Hon., Chapter 9 „Acoustic Modeling" in Spoken Language Processing- A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001.

[20] Huang, X., Acero, A., Wuen-Hon., Chapter 1 „Introduction" in Spoken Language Processing- A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001.

[21] Wikipedia, Prosody, http://en.wikipedia.org/wiki/Prosody_(linguistics), accessed at 30/05/2014. [22] Jurafsky, D., Martin, J., "Automatic Speech Recognition," Chapter 7 "Speech Decoding" in Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition (2nd Ed.), Pearson Education, 2009.

[23] Wikipedia, Mel scale, http://en.wikipedia.org/wiki/Mel_scale, accessed at 31/05/2014 [12] Buzo, A., "Automatic Speech Recognition over Mobile Communication Networks," Teză de doctorat, Universitatea Politehnica din Bucureşti, România,

[24] Wikipedia, MFCC, http://en.wikipedia.org/wiki/MFCC, accessed at 31/05/2014 45.

[25] Huang, X., Acero, A., Wuen-Hon., Chapter 8 „Hidden Markov Models" in Spoken Language Processing- A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001.

[26] Gales, M., Young, S., "The Application of Hidden Markov Models in Speech Recognition" .

[27] Wikipedia, WER, http://en.wikipedia.org/wiki/Word_error_rate accessed at 29/05/2014.

[28] http://www.grymoire.com/unix/sed.html, accessed at 30/06/2014 [18] https://www.ffmpeg.org/, accessed at 30/06/2014.

[29] Wikipedia, http://en.wikipedia.org/wiki/Speaker_diarisation, accessed at 30/06/2014.

[30] Wikipedia, SMT, http://en.wikipedia.org/wiki/Statistical_machine_translation accessed at 10/06/2014, accessed at 03/07/2014.

[31] CMUSphinx Tutorial, Training Acoustic Model, http://cmusphinx.sourceforge.net/wiki/tutorialam, accessed at 10/06/2014.

[32] https://kb.iu.edu/d/afik, accessed at 02/06/2014.

[33] http://www.fileformat.info/tip/linux/iconv.htm, accessed at 03/06/2014.