ITU-T Kaleidoscope 2010
Beyond the Internet? - Innovations for future networks and services

# CROSS-LANGUAGE IDENTIFICATION USING THE WAVELET TRANSFORM AND ARTIFICIAL NEURAL NETWORK

*Shawki A. Al-Dubaee, and Nesar Ahmad*


**Department of Computer Engineering,
Aligarh Muslim University,
Aligarh-202002, U.P, India.**

**shawkialdubaee@gmail.com,
n.ahmad.ce@amu.ac.in**

**Pune, India, 13 – 15 December 2010**

# Outline of Talk

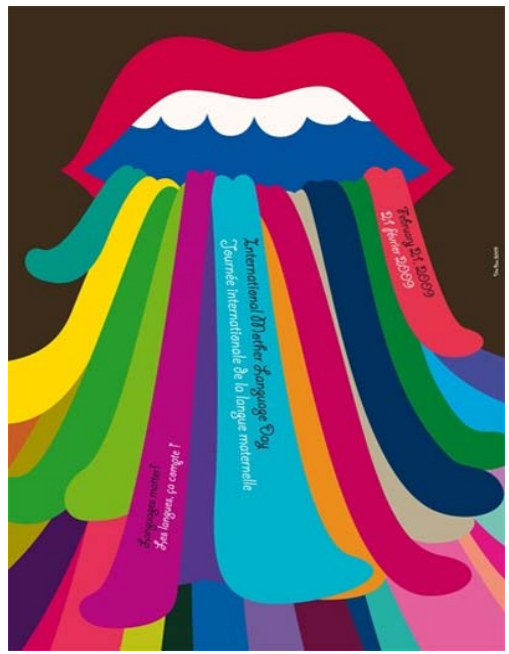
- Overview
- Multilingualism
- Query/Document as signal
- Wavelet Transforms
- Urdu Language
- Database
- The proposed model
- Flowchart for Power Deviation method
- Architecture of ANN and Classification ratio
- Conclusions

# OVERVIEW

- New era of Internet
- Where We Are Today? & Does the digital divide still exit?
  - High Technology revolution.
  - dynamic, heterogenous,unlabeled ,high dimensional and time varying nature of the Web.
  - ” to achieve worldwide access to e-contents in all languages, improve the linguistic capabilities of users and create and develop tools for multilingual access to the Internet." [UNISCO source]

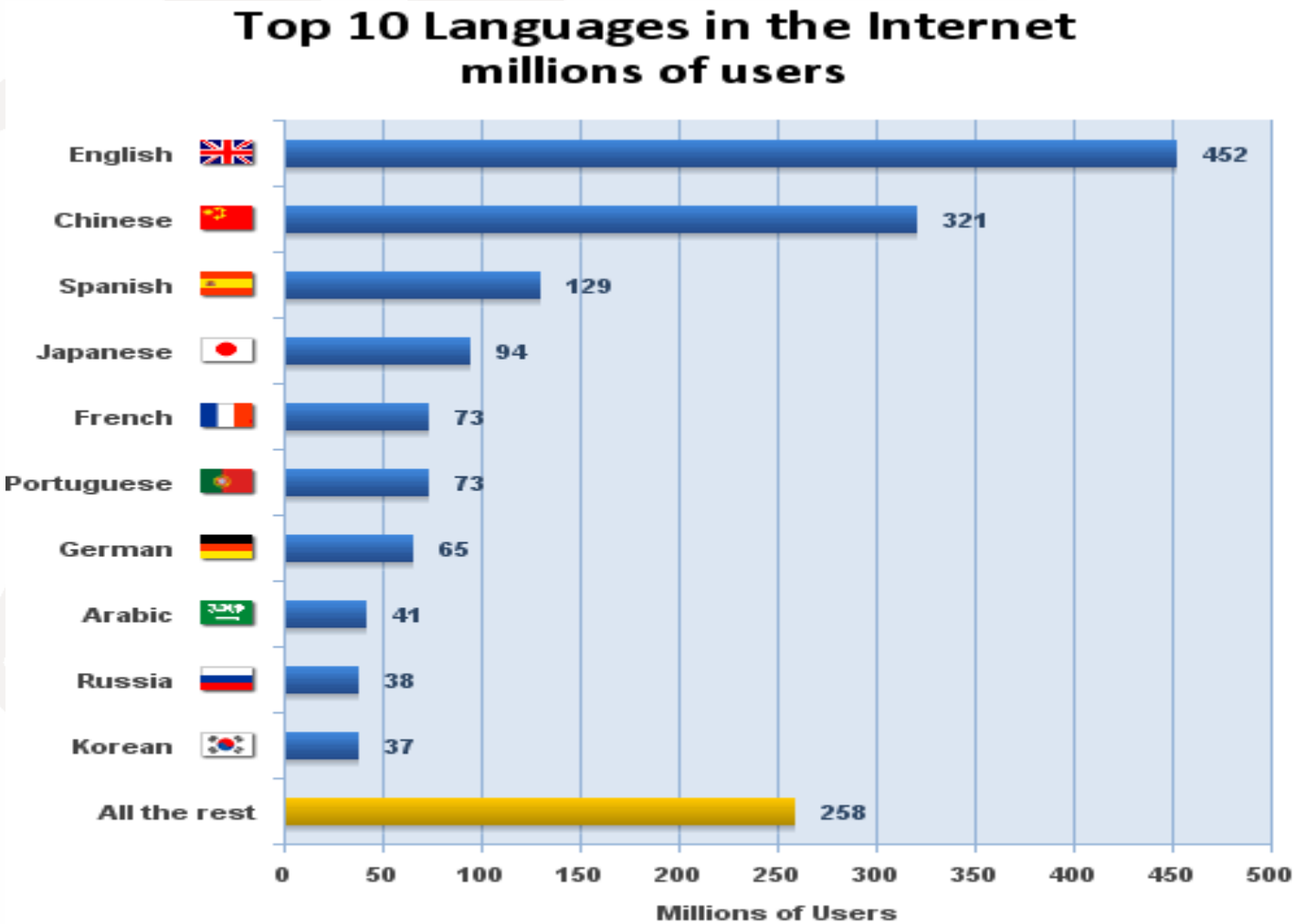
# Multilingualism

- Multilingualism
  - Ability to communicate in a number of languages
- Monolingual
  - Using or knowing only one language.
- Multilingual
  - Having the ability to use several languages.
- Multilingual vs Monolingual




**Mother-Earth (source of UNESCO)**

# Top of 10 Internet Languages




Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated Internet users is 1,581,571,589 for 2008

# Query/Document as Signal

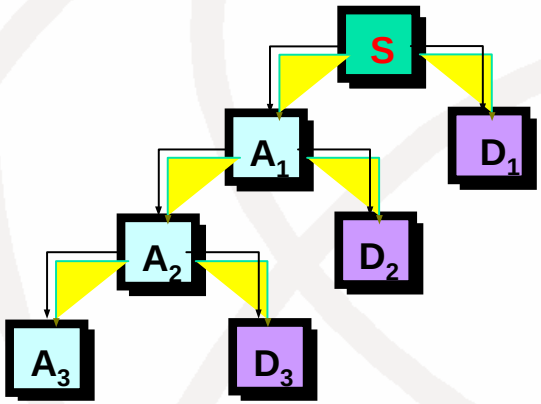
- Why do we need to convert query/document to Signal?
  - To perform many complex computations.
  - To facility applying mathematical transformations, especially the wavelet transform where the signal has been transformed by the availability of mathematical transforms. Therefore, we can obtain further information from the signal that is not readily available in the raw signal or normal language .
  - We can search easily about chemical formals and mathematical/scientific symbols.
  - It may enable to be used parallel processing of query because our method supports framing query or document.
  - To solve the problem of Multilingual Web information retrieval.

- To make any language search queries crafted in one language (Signals), regardless of type, script, word order, direction of writing and difficult font problems of the language.
- To facility worldwide access to the information superhighway using one's own languages of the Internet users.
- These signal queries do not exclude anything from the query (such as the spaces, commas, question marks, and prepositions or applying parsing or stemming on the query). One of the reasons of that, we need to keep the position of every word in the query in order to get full meaning of the query and track the content of document .
- Some of the properties of the wavelet transform enable indexing of large size data with a small number of terms, which in turn facilitate faster and more accurate searches where the computational complexity of wavelet transform is O(n).

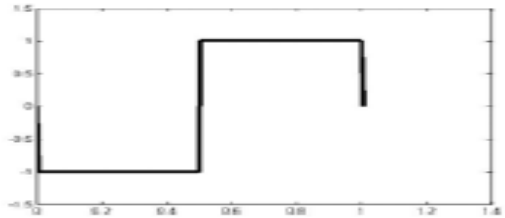
- To preserve endangered languages and cultures.

- We need to adopt a combination of this technique, the wavelet transform, and artificial neural network, to mimic the brain processes of human being in bringing and searching the information required. The brain of human being deals with the query (reading or thinking) as a signal and gets the information required.

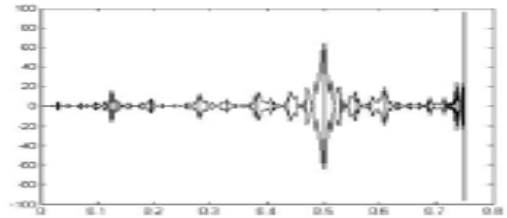
# Wavelet Transforms



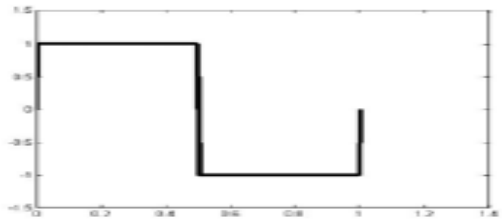

Three levels decomposition of FWT



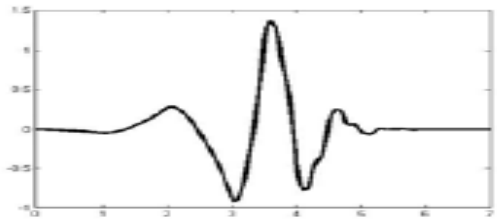
a) The sym1 wavelet function.



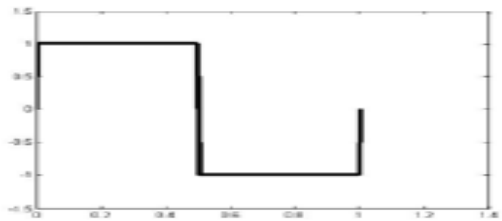
b).The bior 3.1 wavelet function



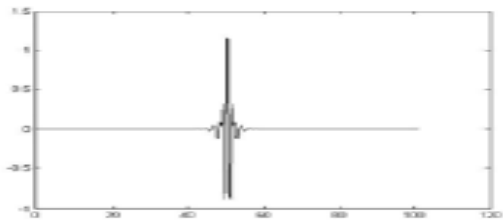
c) The haar wavelet function.



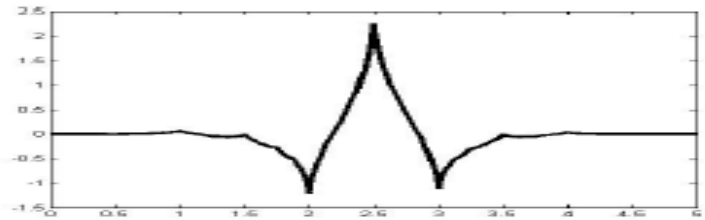
d) The db4 wavelet function.



e) The db1 wavelet function.

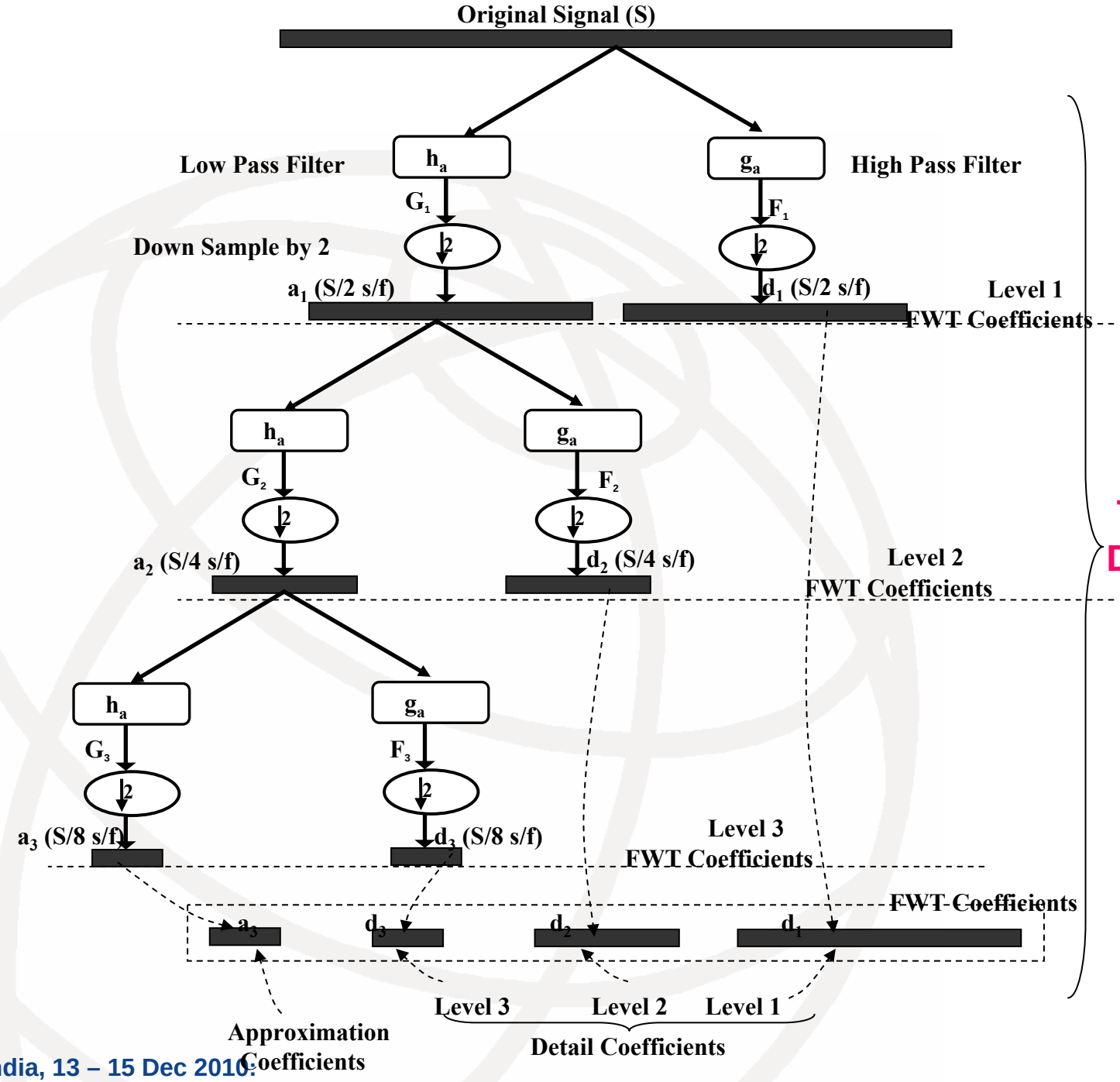


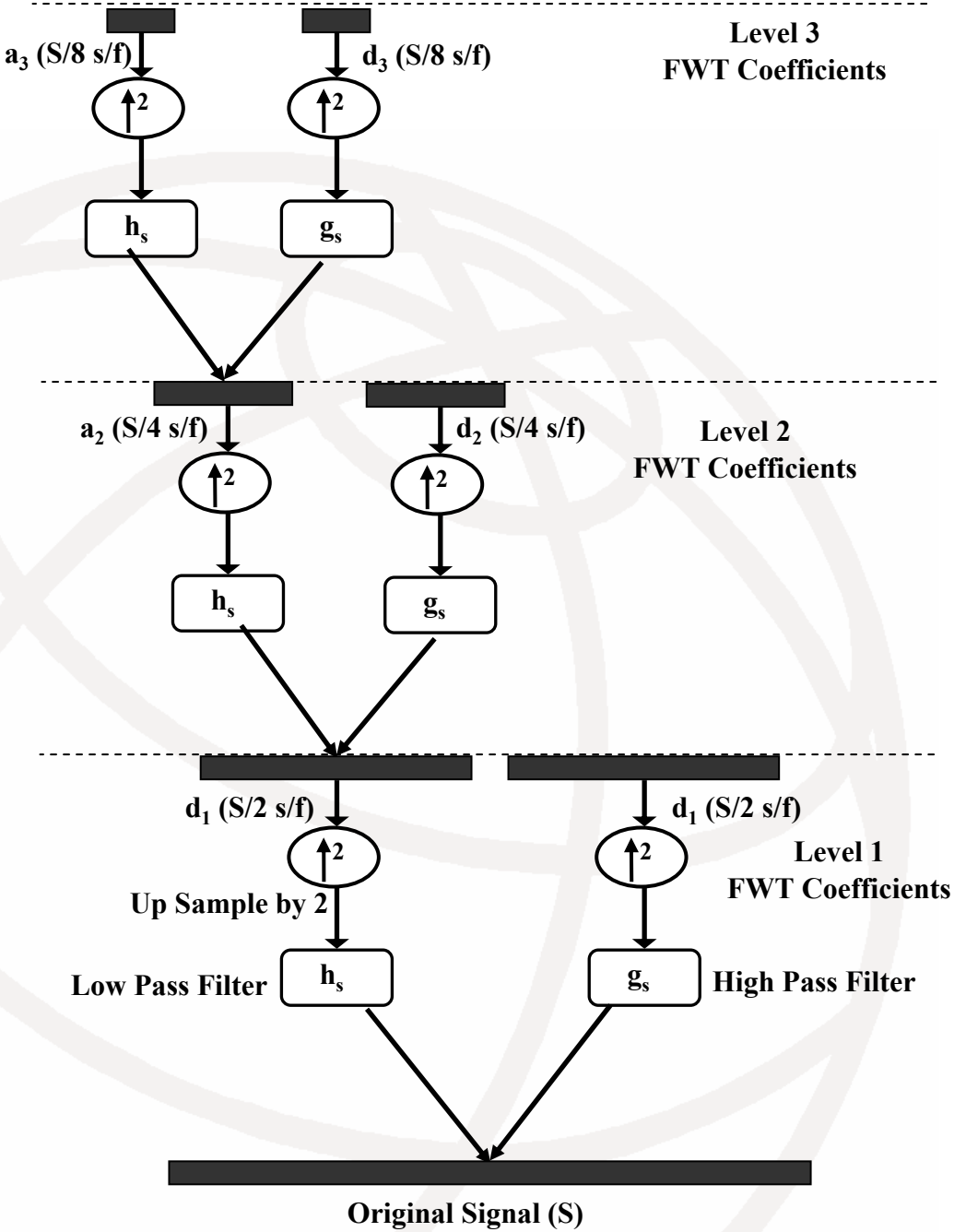
f) The demy wavelet function.



g) The coif 1 wavelet function.

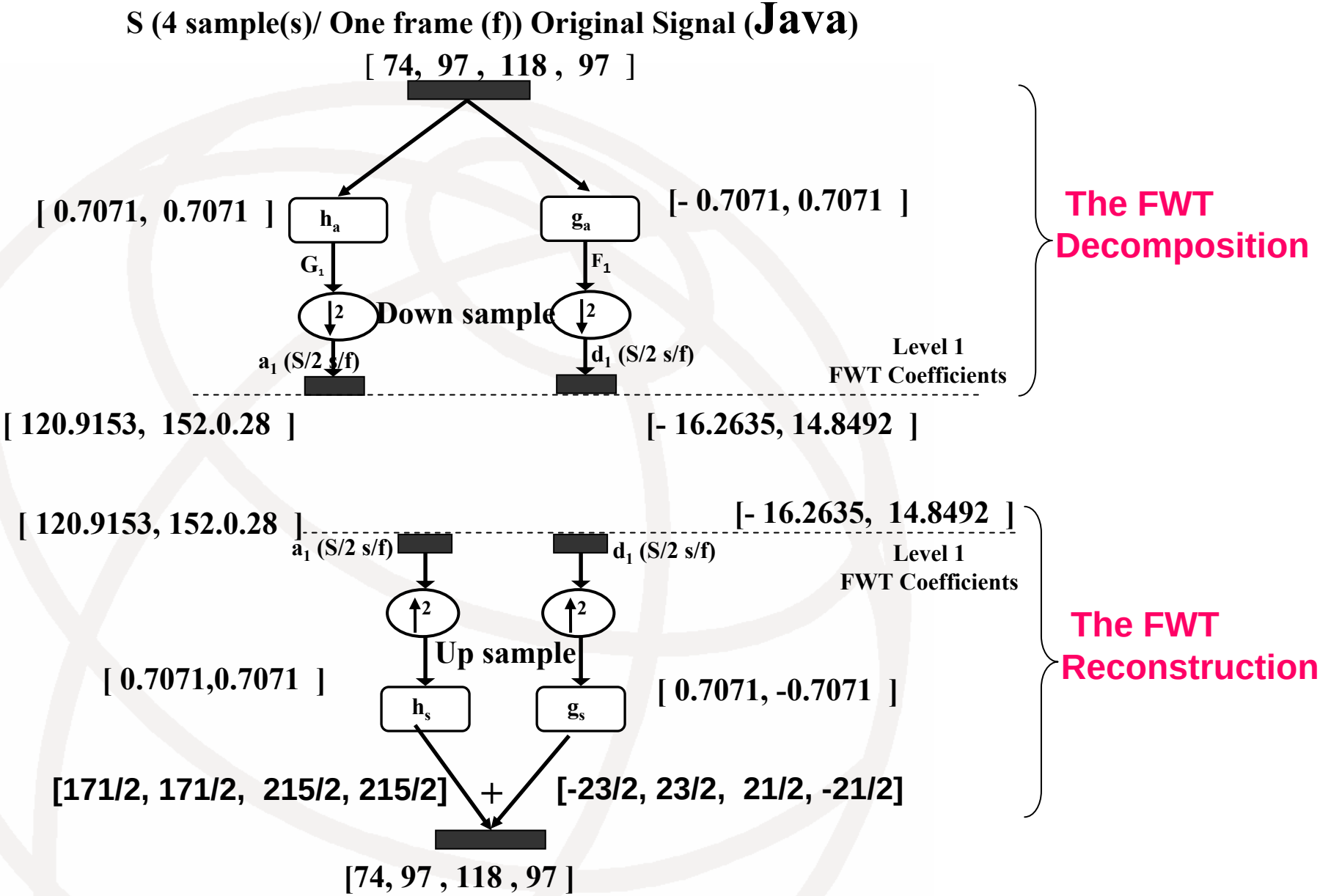
Some of wavelet families


Original Signal (S)
Low Pass Filter
$h_a$
$g_a$
High Pass Filter
$G_1$
$F_1$
Down Sample by 2
2
$a_1$ (S/2 s/f)
$d_1$ (S/2 s/f)
Level 1
FWT Coefficients
$h_a$
$g_a$
$G_2$
$F_2$
2
2
$a_2$ (S/4 s/f)
$d_2$ (S/4 s/f)
Level 2
FWT Coefficients
$h_a$
$g_a$
$G_3$
$F_3$
2
2
$a_3$ (S/8 s/f)
$d_3$ (S/8 s/f)
Level 3
FWT Coefficients
FWT Coefficients
$a_3$
$d_3$
$d_2$
$d_1$
Level 3
Level 2
Level 1
Approximation Coefficients
Detail Coefficients
The FWT Decomposition




The FWT Reconstruction

# Demonstration Example



| wavelets | Reconstruction question query of "What is your name?" for English language. | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Level 1 | % | Level 2 | % | Level 3 | % | Level 4 | % |
| Haar | What-is-your-name? | 83 | What-is-your-name? | 83 | What-is-your name? | 89 | What is your name? | 100 |
| Db 1 | What-is-your-name? | 83 | What-is-your-name? | 83 | What-is-your name? | 89 | What is your name? | 100 |
| Db2 | Wg`s hr xntq n`ld? | 33 | Wg`s hr xntq n`ld? | 33 | Wg`s ir xntq n`le? | 44 | Wg`s hr xntq n`le? | 39 |
| Db 3 | Wh`t-hs-ynur-nale> | 56 | Wg`t-hr-your-name> | 56 | Wg`t-ir-your-m`me> | 50 | Wg`t-hr-your-m`me> | 44 |
| Db 4 | Vgas ir xntq m`ld? | 33 | Vgas hr xntq mald? | 33 | Vg`s hr xntq m`ld? | 22 | Vg`s hr xntq mald? | 28 |
| Db 5 | Wh`t-is-your-n`me> | 67 | Vhat-is-your-name> | 72 | What-hs-your-namd> | 67 | Vhat-hs-your-namd> | 61 |
| Db 6 | Vh`t-hs-ynur-n`ld? | 44 | Vh`t-hs-ynur-n`ld? | 44 | Wh`s-hs-ynur-n`ld? | 44 | Wh`s-hs-ynur-n`ld? | 44 |
| Db 7 | Vhat-is-your-name> | 72 | Vhat-hs-your-mame> | 61 | Vhat-hs-your-mame> | 61 | Vhat-hs-your-name> | 67 |
| Db 8 | Wh`t-hs-your-n`me> | 61 | Wh`t-hs-your-m`me> | 56 | Wh`t-hr-your-m`md> | 44 | Wh`t-hr-your-m`md> | 44 |
| Db 9 | Wg`t hr yntr n`ld? | 50 | Wg`t hr yntr n`ld? | 50 | Wg`t hr yntr n`ld? | 50 | Wg`t hr xntr n`ld? | 44 |
| Db 10 | Wg`s hr xntq m`md? | 33 | Wg`s hr xntq m`ld? | 28 | Vg`s hr xntq m`ld? | 22 | Vg`s hr xntq m`ld? | 22 |
| Coif 1 | Vhas-is-xouq-mame> | 50 | Vhas-is-youq-mame> | 56 | Vhas-is-youq-mame> | 56 | Vhas-is-your-mame> | 61 |
| Coif 2 | Wg`s hr xntq m`ld? | 28 | Vg`s ir xntq n`ld? | 33 | Vg`s is xntq m`ld? | 33 | Vg`s is xntq m`ld? | 33 |
| Coif 3 | Wgat ir-yotq-mamd> | 50 | Wgat-hr-yotq namd> | 50 | Wgat-hr-yotq namd> | 50 | Wgat-hr-yotq namd> | 50 |
| Coif 4 | Vhas-is-youq-nald? | 56 | Vgat is-xotq-nald? | 50 | Vgat is-xotq-nald? | 50 | Vgat is-xotq-nald? | 50 |
| Coif 5 | Wh`s is-xouq-nald? | 56 | Vh`s is-xouq-nald? | 50 | Vhat is-xouq-nald? | 61 | Vhat is-xouq-nald? | 61 |
| Sym 1 | What-is-your-name? | 83 | What-is-your-name? | 83 | What-is-your name? | 89 | What is your name? | 100 |
| Sym 2 | Wg`s hr xntq n`ld? | 33 | Wg`s hr xntq n`ld? | 33 | Wg`s ir xntq n`le? | 44 | Wg`s hr xntq n`le? | 39 |
| Sym 3 | Wh`t-hs-ynur-nale> | 56 | Wg`t-hr-your-name> | 56 | Wg`t-ir-your-m`me> | 50 | Wg`t-hr-your-m`me> | 44 |
| Sym 4 | Vhas-is-xouq-mame> | 50 | Vhas-is-youq-mame> | 56 | Vhas-is-youq-mame> | 56 | Vhas-hs-youq-mame> | 50 |
| Sym 5 | Wg`t hr yntr n`ld? | 50 | Wg`s hr yntr n`ld? | 44 | Wg`t ir xntr n`ld? | 50 | Wg`t ir yntr n`ld? | 56 |
| Sym 6 | Vhas-is-xouq-mame> | 50 | Vhas-is-xouq-mame> | 50 | Vhas-is-youq-mame> | 56 | Vhas-is-youq-mame> | 56 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Sym 7 | Wg`t hr yntr n`ld? | 50 | Wg`t hr yntr n`ld? | 50 | Wg`t hr yntr n`ld? | 50 | Wg`t hr yntr n`ld? | 50 |
| Sym 8 | Vhas is xouq mamd? | 67 | Vhas is xouq m`md? | 61 | Vhas is xouq m`md? | 61 | Vhas is xouq m`md? | 61 |
| Sym 9 | Vgas-is-xouq-mamd> | 39 | Vgas-hr-xouq-mald> | 22 | Vgas-hr-xouq-mald> | 22 | Vgas-hr-xouq-mald> | 22 |
| Sym 10 | Vhat-is-your-name> | 72 | Vhat-is-your-mame> | 67 | What-hs-your-mame> | 67 | What-hs-your-mamd> | 61 |
| Bior1.1 | What-is-your-name? | 83 | What-is-your-name? | 83 | What-is-your name? | 89 | What is your name? | 100 |
| Bior1.3 | What is your-name? | 94 | What is your-name? | 94 | What is your-name? | 94 | What is your name? | 100 |
| Bior1.5 | What-is your name? | 94 | What-is your name? | 94 | What is your name? | 100 | What is your name? | 100 |
| Bior2.2 | What is your name? | 100 | What is your name? | 100 | What is your name? | 100 | What is your name? | 100 |
| Bior2.4 | What-is your name? | 94 | What-is your name? | 94 | What is your name? | 100 | What is your name? | 100 |
| Bior2.6 | Wh`t is your-namd? | 83 | What is your name? | 100 | What is your name? | 100 | What is your name? | 100 |
| Bior2.8 | What is-your name? | 94 | What is your name? | 100 | What is your name? | 100 | What is your name? | 100 |
| Bior3.1 | What is your name? | 100 | What is your name? | 100 | What is your name? | 100 | What is your name? | 100 |
| Bior3.3 | What is-your-name? | 89 | What is your-name? | 94 | What is your name? | 100 | What is your name? | 100 |
| Bior3.5 | What is-your name? | 94 | What is your name? | 100 | What is your name? | 100 | What is your name? | 100 |
| Bior3.7 | What hr-your name? | 83 | What is-your-name? | 89 | What is your name? | 100 | What is your name? | 100 |
| Bior3.9 | What is your name? | 100 | What is your name? | 100 | What is your name? | 100 | What is your name? | 100 |
| Bior4.4 | Wg`s hr xntq m`ld? | 28 | Vg`s ir xntq n`ld? | 33 | Vg`s ir xntq n`ld? | 33 | Vg`s ir xntq n`ld? | 33 |
| Bior5.5 | Wh`t-is-your-name> | 72 | Wh`t-ir-your-name> | 67 | Wh`t-hr-your-namd> | 56 | Wh`t-hr-your-name> | 61 |
| Bior6.8 | Wg`s hr xntq m`ld? | 28 | Vg`s ir xntq n`ld? | 33 | Vg`s ir xntq n`ld? | 33 | Vg`s ir xntq m`ld? | 28 |
| Dmey | Vhat-hs-yotr-mamd> | 50 | Vhat-hs-your-mamd> | 56 | Vhat-hs-your-mamd> | 56 | Vhat-hs-your-mamd> | 56 |
| average | | 64 | | 64 | | 65 | | 65 |

# Urdu Language

•Urdu language is a Central Indo-Aryan language of the Indo-Iranian branch which belongs to the Indo-European family of languages.

• There are about 104 million Urdu language speakers, including those who speak it as a second language (http://www.omniglot.com/writing/urdu.htm).

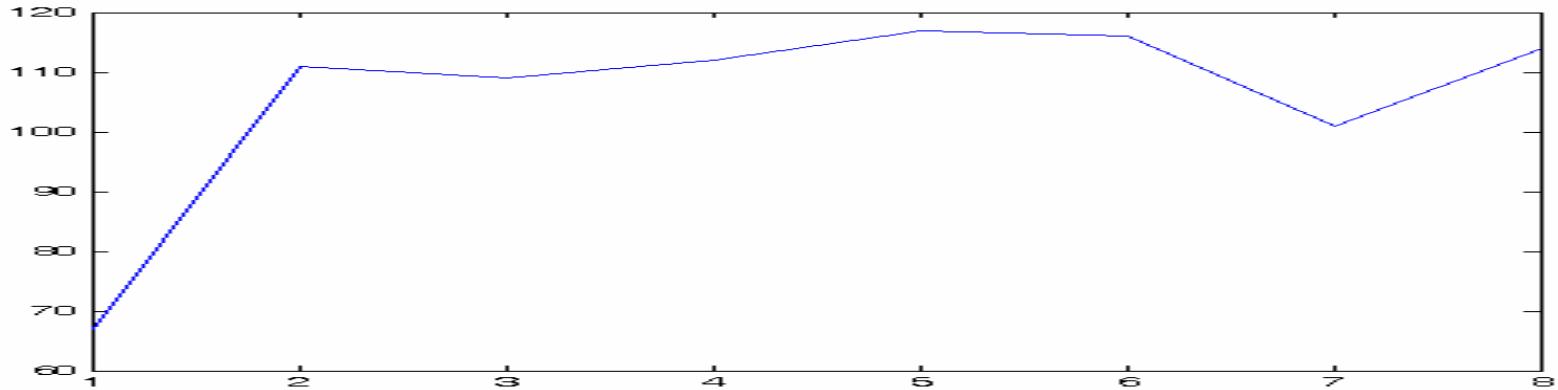
• Urdu scripts and numbers are written from right to left and left to right respectively. Its scripts have shapes similar to Arabic, Persian language characters.


| | | | | | |
|---|---|---|---|---|---|
| ث | ٹ | ت | پ | ب | ا |
| ڈ | د | خ | ح | چ | ج |
| س | ژ | ز | ڑ | ر | ذ |
| ع | ظ | ط | ض | ص | ش |
| ل | گ | ک | ق | ف | غ |
| ء | ھ | ہ | و | ن | م |
| | | | | ے | ی |

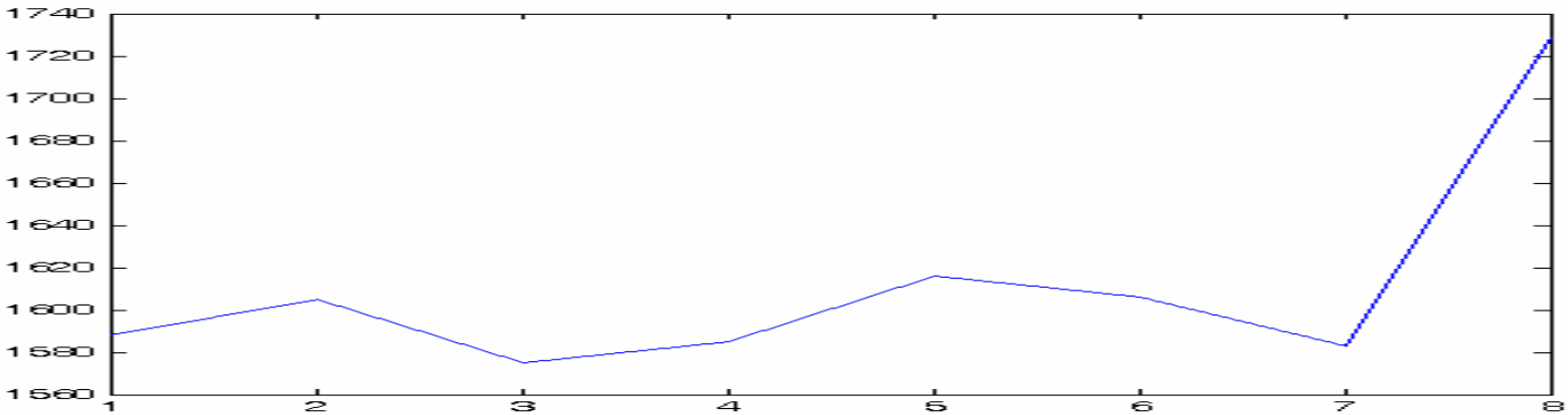
# Query to signal (Database)

| No. | English Queries | Urdu Queries |
|---|---|---|
| 1 | Computer | شمارندہ |
| 2 | Computers cannot "think" for themselves in the sense that they only solve problems in exactly the way they are programmed to, and arrive at the correct answer (500,500) with little work. In other words, a computer programmed to add up the numbers one by one as in the example above would do exactly that without regard to efficiency or alternative solutions. | اور اس متبادل راہ کے استعمال سے انسان وہی درست جواب (500500) نکال لیتا ہے جو شمارندہ اوپر دی گئی ہدایات سے نکالے گا۔ بس یہ فرق (سوچنے کا) شمارندے اور انسان میں ایسا ہے کہ جس کی بنا پر شمارندے مکمل خود مختار نہیں ہوتے۔ |
| 3 | Punch Card | کا ایک سوراخی بطاقہ |
| 4 | What's your name? | آپ کانام کیاہے؟ |
| 5 | Where do you live? | آپکا تعلق کہاں سے ہے؟ |

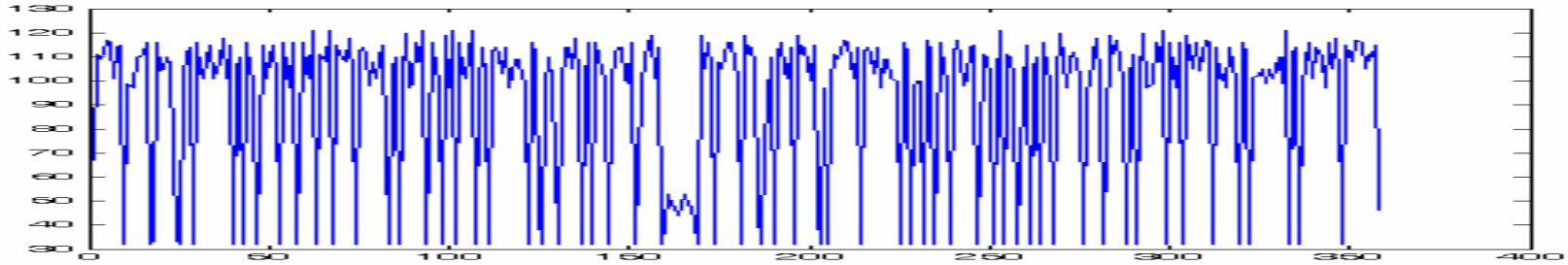
# Some of signal queries (a to f ) for the English and Urdu languages.



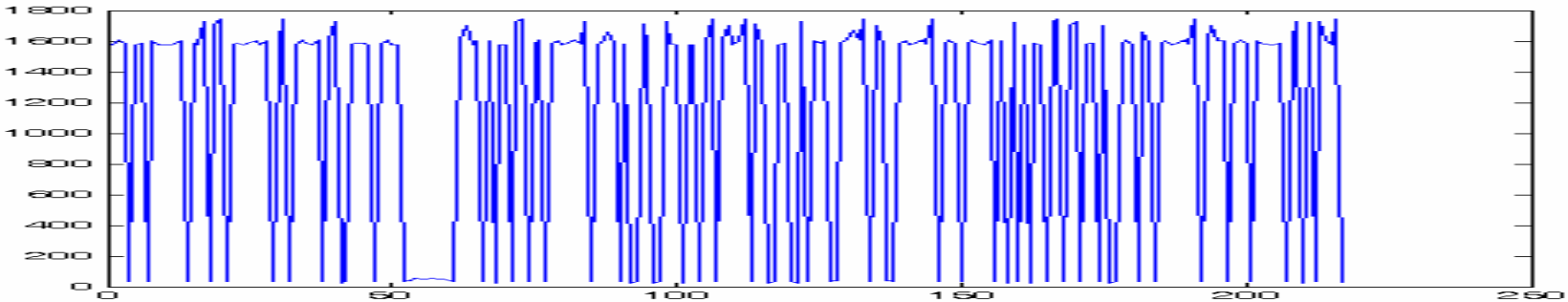

a) English query No.1.(Computer )



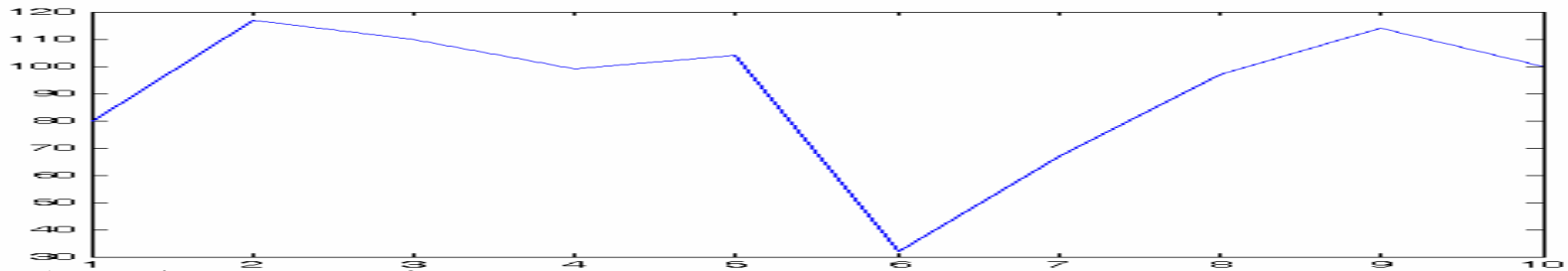

b) Urdu query No.1. (شمارندہ )



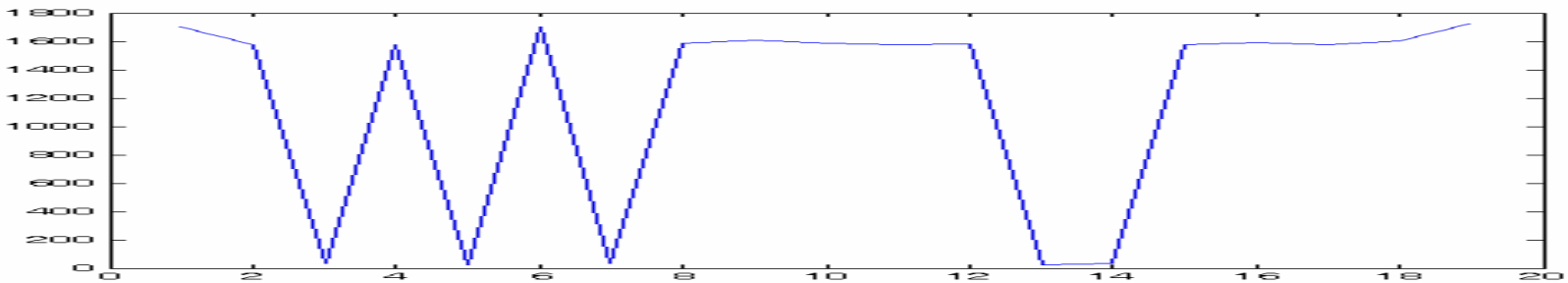
c) English query No.2.



d) Urdu query No.2.

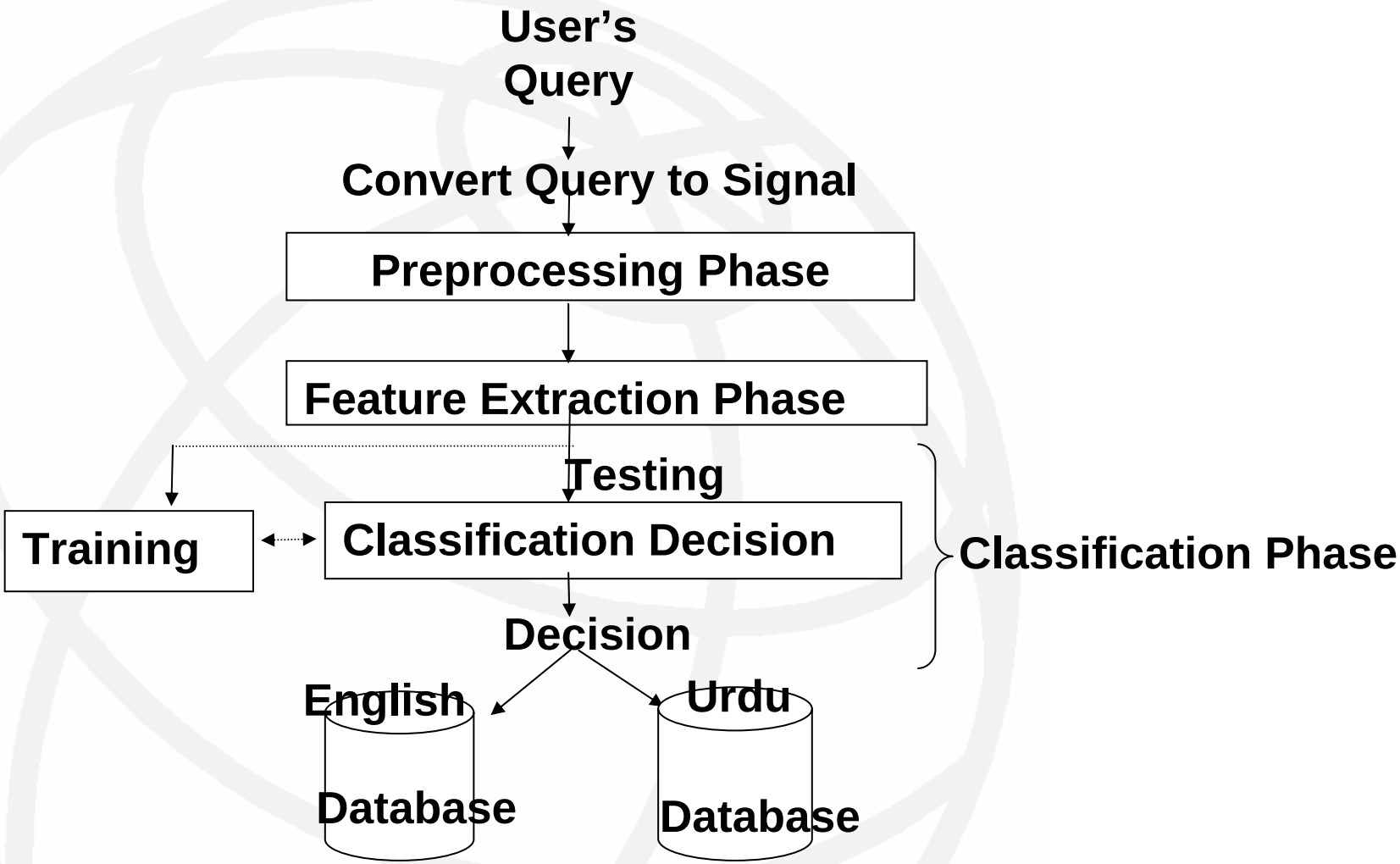


e) English query No.3.

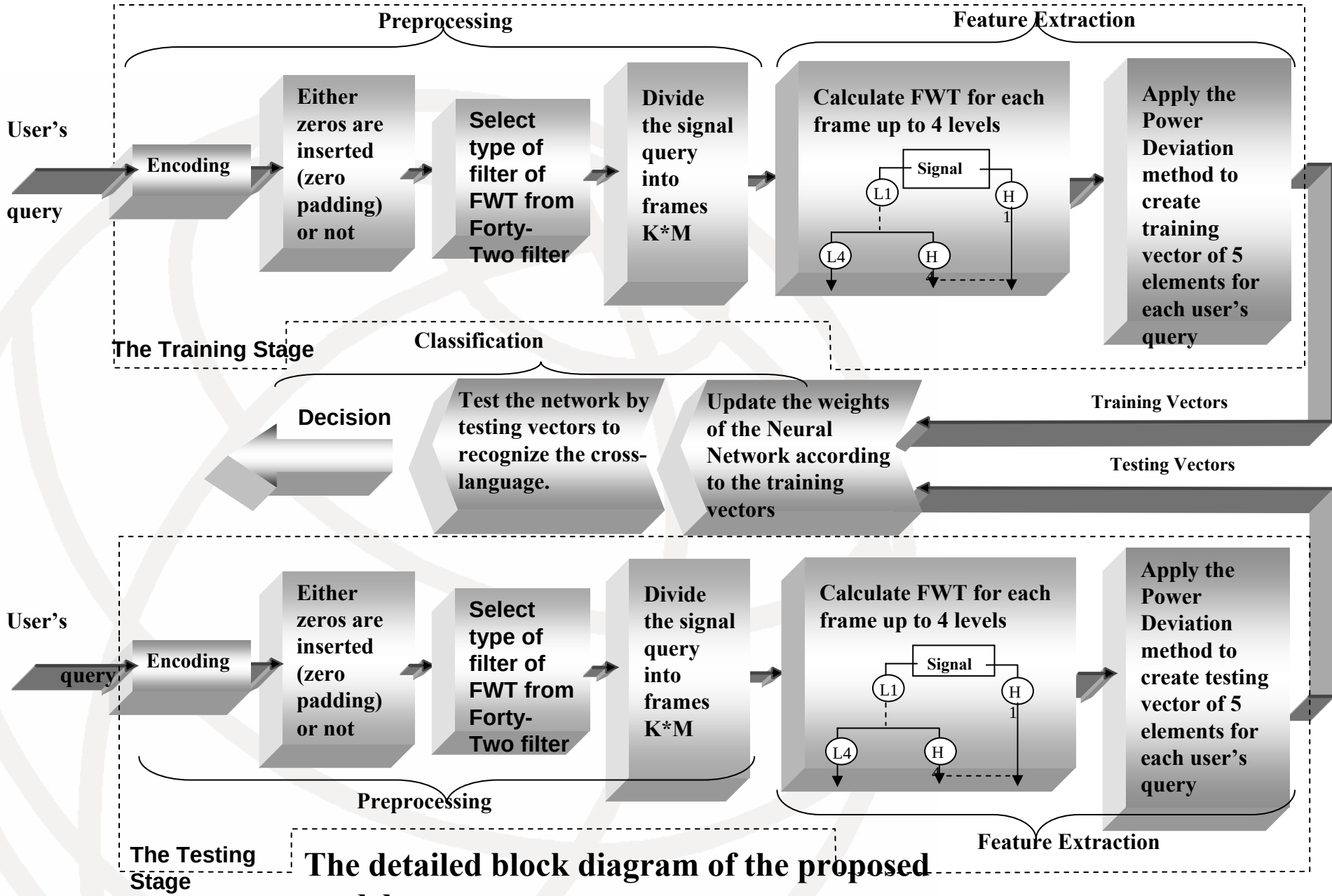


f) Urdu query No.3.

# The proposed Model

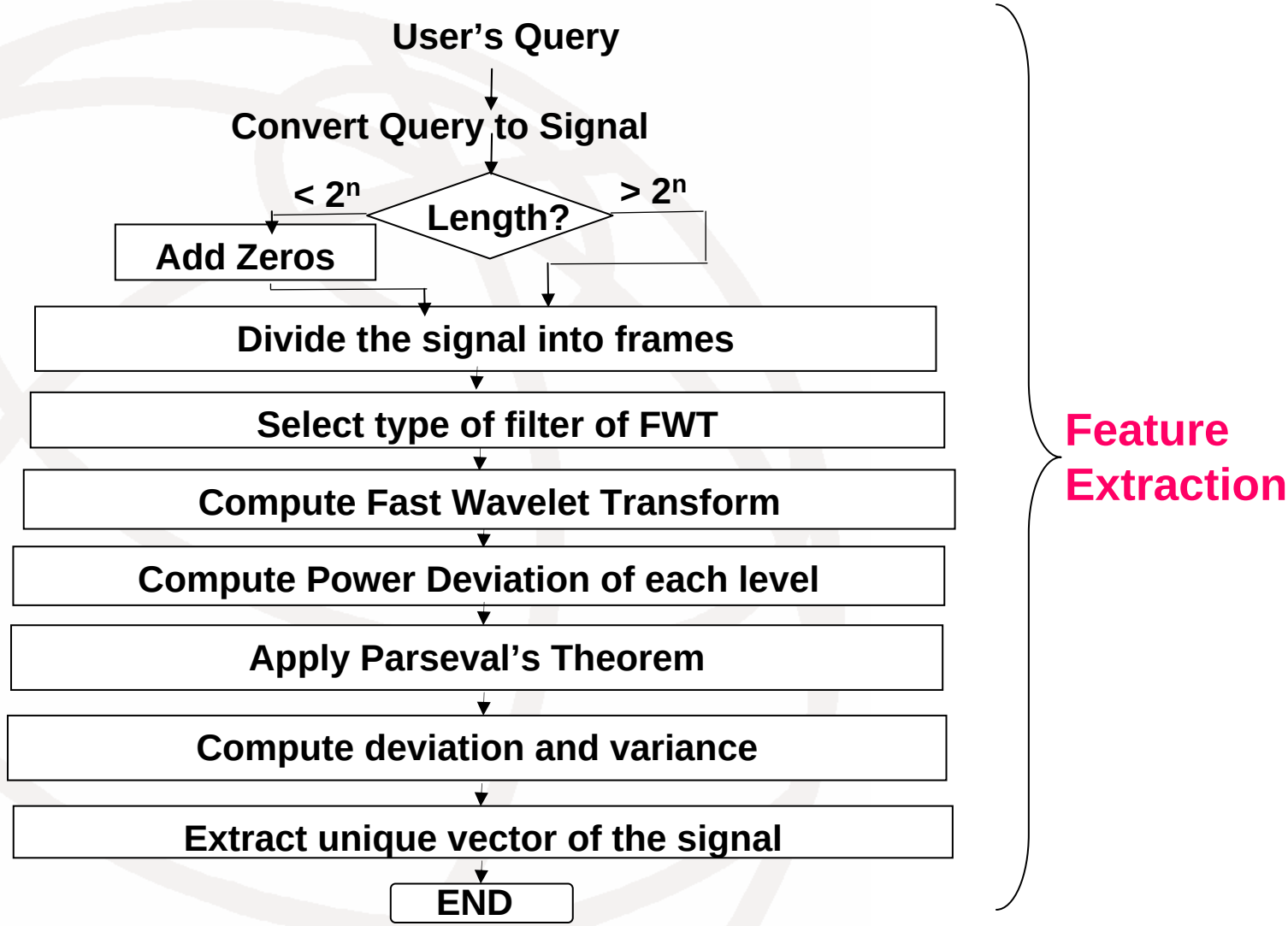



Methodology for classification of the processed signal.



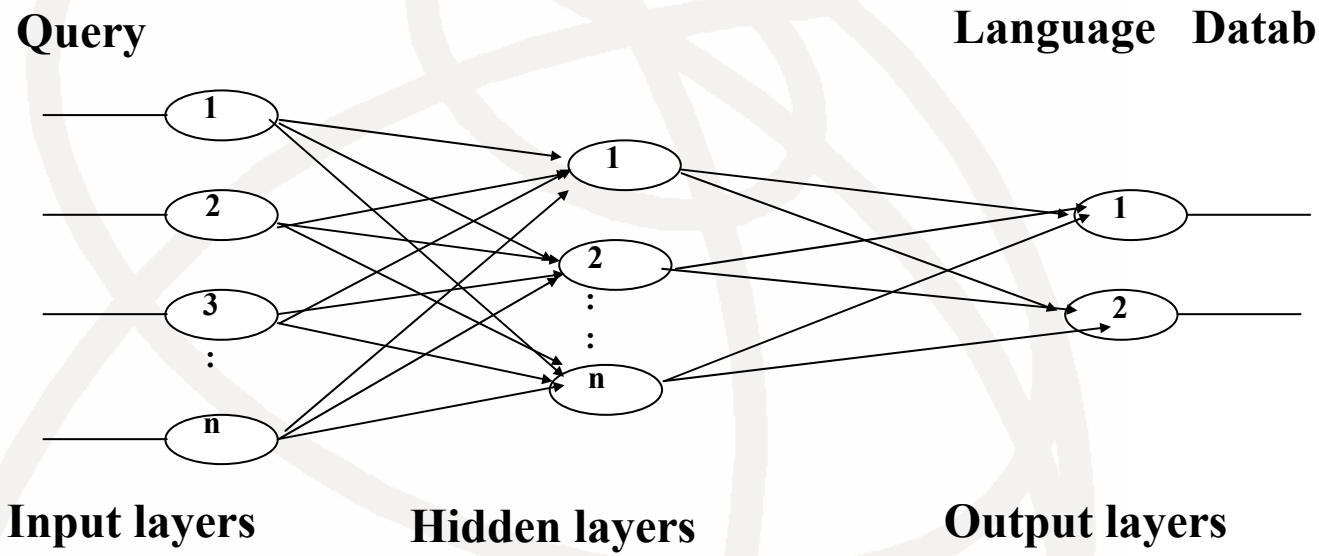

**The detailed block diagram of the proposed model**

# Flowchart for Power Deviation method

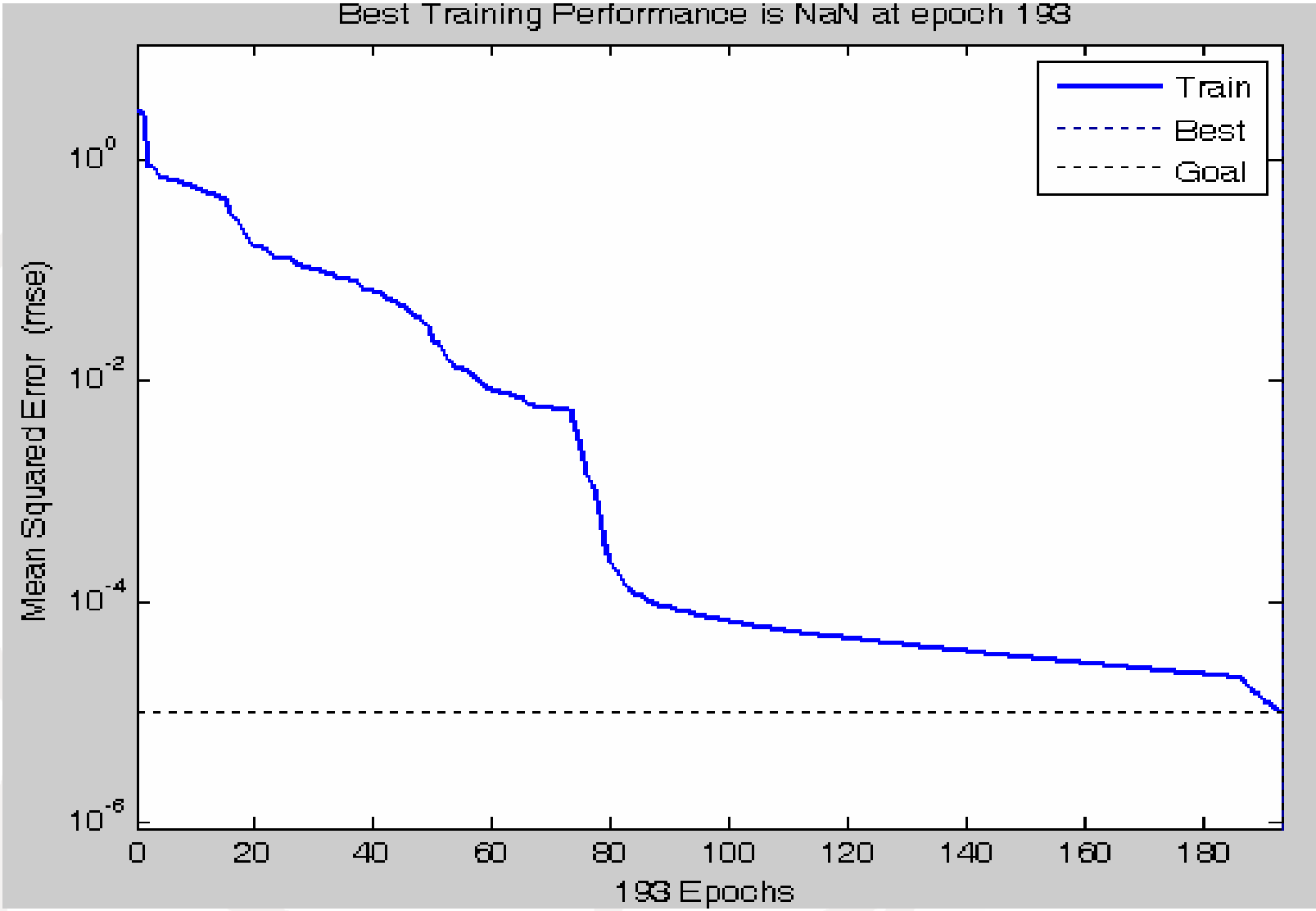



**Flowchart for Power Deviation method.**

# Architecture of ANN and Classification ratio




**The proposal architecture 5-15-2 using ANN of PD method for the English and Urdu languages.**



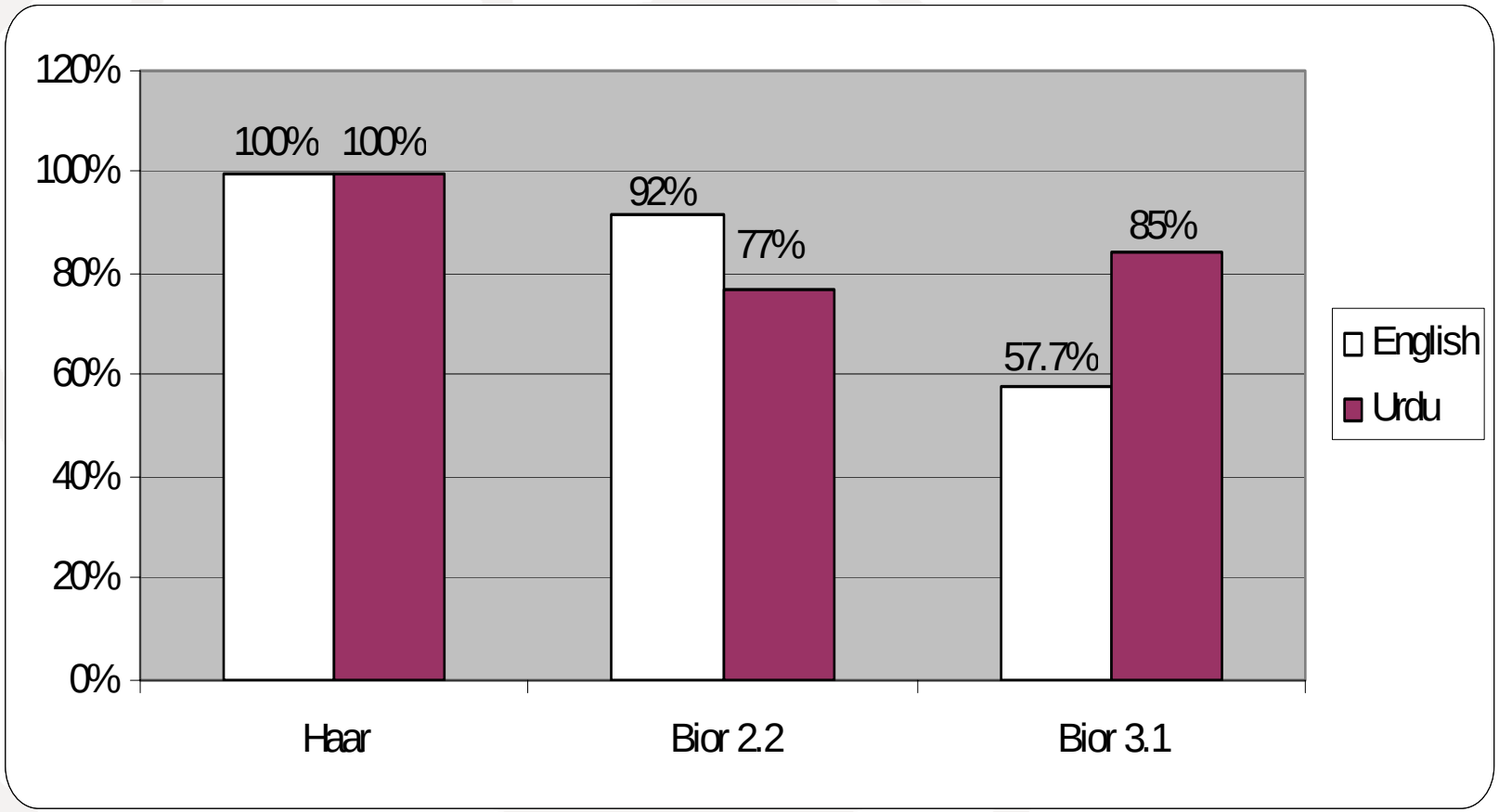

**The training of the ANN dedicated to FWT using (Haar filter with 4.3 seconds) for the English and Urdu languages**

# Classification ratio of ANN

## Table for training Time and classification ratio of ANN.

| FILTERS | TRAINING TIME (SEC) | ACCURCY % | | AVERAGE % |
|---|---|---|---|---|
| | | ENGLISH | URDU | |
| HAAR | 4.3 | 100% | 100% | 100% |
| BIOR 2.2 | 18.1 | 92.3% | 76.9% | 84.6% |
| BIOR 3.1 | 11.6 | 57.7% | 84.6% | 65.4% |

**A typical bar graph plot corresponding to the percentage of correct identification for the English-Urdu cross-languages using Haar with 4.3 seconds.**



# Conclusions

• We propose a novel approach for feature extraction and classification of signal queries. This novel approach presents an automatic method for classification of English and Urdu languages identification. The classifier used is a three-layered feed-forward artificial neural network which is called Back Propagation and the feature vector is formed by calculating the wavelet coefficients.

• The best performance of language identification of English and Urdu languages is the Haar filter with 100% as classification ratio and good training time 4.3 seconds.

•We need to adopt a combination of this technique, the wavelet transform, and artificial neural network, in order to mimic the brain processes of human being in bringing and searching the information required. The brain of human being deals with the query (reading or thinking) as a signal and gets the information required.


Thank You