CHAPTER 4

RESEARCH METHODOLOGY

Chapter Overview:

- Introduction to Research Methodology
- Speech Corpus Development
- Selection and Normalization of Text
- Software use for Recording, Formant Frequency and recording Duration
- Setup for speech Recording
- Speaker Selection Criterias
- Procedure of Recording
- Database Creation Procedure
- Methodology used for Pitch, Formant Frequency, MFCC, LPC, Short Time Energy, and Zero Cross Rate Analysis
- Description of Tools Used For Recording / Synthesis
- Selected Phonemes
- Sample Selection
- Noise Filtering of Speech Data
- Word Selection
- Programming Software Tool-Matlab R2009a
- Database creation, Recording and Organization
- Annotation
- Tagging Policy with Typical Examples
- Intonation in Bodo Language
- Database Management Structure

CHAPTER 4

RESEARCH METHODOLOGY

4.1 Introduction:

To increase the knowledge in a particular field, the formal work done systematically using scientific methods which includes facts and knowledge of humanity, culture and society can be termed as Research and Experimental development activity. The knowledge acquired is used further for new developments. In the field of speech research too same approach is followed to achieve newer goals. Research methodology is defined as a process of searching/finding a solution to any research task [194, 195]. So, it is a science of understanding the scientific procedure/strategy adopted during this research. It is time to discuss the strategies and logic involved herewith that are in general followed. Researchers must design a methodology for the problem chosen for his/her work is a basic and very essential activity. Moreover methods applied in two separate tasks may be same but methodology may not be same. Research activity requires skillfully performed careful investigation or analysis in search of new fact finding in any branch of knowledge. Aim of the speech Technology research fraternity principal goal is to investigate targeting fresh objectives in this field. Research and experimental development is the formally executed set of tasks performed systematically to prosper knowledge to higher level, may be knowledge of humanity, culture and society. *It* is the way of finding systematical solution of any research problem. We can understand it as a science of studying how research is done systematically. With above discussed points in view, the following steps are adopted in our work.

Definition: Research is a scientific inquiry aimed at learning new facts, testing ideas, etc. It is the systematic collection, analysis and interpretation of data to generate new knowledge and answer a certain question or solve a problem.

4.2 Speech Corpus Development:

The final rate of authentication is heavily affected by the quality of the acoustic unit inventory used as source of object in the experiments. There are many factors that contribute to the quality of the acoustic unit inventory:

i. The source voice corpus that provides test units for experiment.
ii. Accuracy achieved while labeling the utterances
iii. Instances present in individual test voice unit
iv. Prosodic richness of all units selected for experiment etc.

Speech corpus is prepared following the steps: Pre-processing, Setting rules and selecting speech, Voice sample collection from speakers (recording), and Orthographic annotation (requires revision occasionally)

4.2.1 Text Sample Choosing and Normalization:

To select phonetically rich text data sources utilized in this research work are:

- ✓ Books written in Bodo, Suitable articles from Magazines/Newspapers, Articles/lectures by established personalities working in these sector, academicians etc.

We ensure that text data include samples from all sector of life Local culture, ethics, literature, livelihood, education, socio-economic condition etc. In this research work approximately 2500 sentences are considered. These data are verified by experts for speech research suitability. Using UTF8, texts are saved for next level of processing.

4.2.2 Recording software applied:

Wave Surfer 1.8.5 version is deployed setting the attributes as follows: Resolution is set at 16bit PCM (Mono) in 16 KHz Sampling Frequency. Then 5 male & female informants record their voice for 8 hours each approximately. UTF-8 text format used to keep text in database.

4.2.3 Speech acquisition environment:

- Behringer C-1U, USB Studio condenser (16mm capsule) microphone.
- Frequency starts from 40 Hz and goes up to 20 KHz.
- Driver: Realtek High Definition Audio.
- Sound Card: Sound blaster live 5.1, creative sound card.
- 8 inch gap is maintained from the microphone to the informant.
- Studio environment is kept semi anechoic & noise proof.
- Place: Deptt. of Instrumentation & USIC laboratory studio, Gauhati University.

. The Block diagram of the recording setup is shown the Figure 4.1 below:

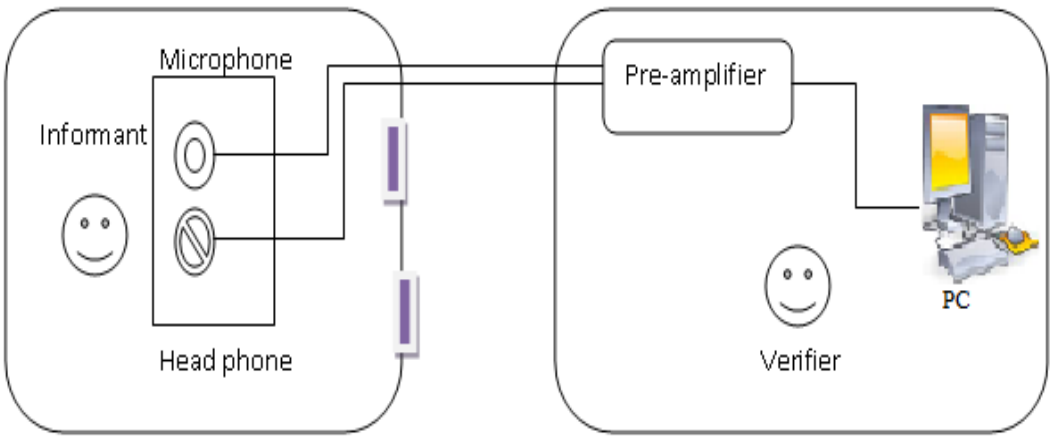

Figure 4.1: Block diagram of arrangements in the studio for data recording.

4.3 Rules for Informant choosing:

Below mentioned factors were looked after before finalizing an informant to acquire voice:

Age group: preference is given to the young generation (from 25 to 35 years)

Spoken language: Only native Bodo speaker are considered.

Qualification: Post Graduation students.

Arrangements were done in the laboratory for an expert (acoustic & phonetic) in Bodo language to monitor the procedure so that richness (Phonetic/Prosodic with consistency and accuracy) of the data can be maintained.

4.3.1 Recording environment description:

After text selection recording begins as follows-

- Place: Dept. of Instrumentation and USIC laboratory studio, Gauhati University.
- Material: printed text provided to all.
- Informants training: they are rehearsed for normal speaking mode.
- Audio data acquisition: After finishing audio recording of sentences, target test voice samples (words) are cut out of it and saved in database using .wav format.

4.3.2 Database Creation Procedure:

In a bid to test and train SPERIA B, database is created with 5 male and 5 female informants. Our selected list includes *one thousand* BODO words having syllable ranging from *one* to *five*. Speaker selection is done such that entire Bodo speaker community is represented and regional balance is maintained. From entire BODO DOMINATED areas of Assam namely, Kokrajhar, Udalguri, Bongaigaon, Baksa

and Kamrup area where variations in spoken dialect is clearly visible are looked upon for this purpose. Above mentioned 5 regions are represented through by at least 1 male and 1 female speakers. Informants are post graduate habitual Bodo speakers belonging to 25 to 35 years of age. Number of utterance for each isolated word for each speaker is kept as 10 times. All together acquisition of *10,000 (ten thousands)* speech samples (utterances) are done and put in the database. Subsequently 60 % samples for training phase and the remaining 40% of are used in the testing phase. Final statistics of the data comes like:

10 informants X 1000 words/informant X 10 samples for each informant/word

=1 Lakh samples.

In the department of Instrumentation & USIC, under the supervision of Dr. Pran Hari Talukdar, a large database has been created and same is used with due permission in this research work. Contents of this database will consist of the following materials and formats: Audio version of test voice samples saved in a single directory applying *.wav* format as shown below: example- "s1f25_05.wav" represents voice sample 1 of female informant of age 25 uttering 5^{th} repetition. Thus the arrangement is:

(Sample index)Male-female (informant index) age (age index) _ (repetition number).wav

4.4 Methodology used for pitch, formant frequency, mfcc, lpc, short time energy and zero cross rate analysis:

In this thesis, analysis of sample test data set is performed for the following features as per the specifications mentioned below:

Level 1: Spectral feature sets: *i. Pitch and ii. Formant frequency*

Level 2: Acoustic feature set: *i. MFCC ii. LPC*

Level 3: Prosodic feature set: *i. STE ii. ZCR*

Vowels and consonants of Bodo language and word set of the Consonant-Vowel (CV), Consonant-Vowel- Consonant (CVC), and Vowel- Consonant (VC) structure.

Employing Wave Surfer 1.8.5 and MATLAB 7.10 we have completed the analysis task. Now for feature study, each digitized voice samples, are framed into 50 frames of 20 millisecond length. Pitch, Formant, MFCC, LPC, STE and ZCR are studied extensively and a set of feature vectors are estimated. In our investigation speech samples of consonant-vowel (CV), consonant-vowel-consonant (CVC) and vowel-consonant (VC) type are considered.

4.5 Description of Tools Used For voice sample acquisition and analysis:

4.5.1 Cool Edit Pro:

This multi track sound editing windows platform software enables us to perform: Filtration using Digital Signal Processing Effect, Multi tracking maximum 64 tracks at a time, Plug-ins extension, Batch processing, Export/import MP3, WAV, WMA files etc.

Voice samples acquisition from informants are done applying cool edit pro 2.0 then keeping, 16 KHz sampling rate it is digitized in .WAV file format. Separately text file are maintained in UTF format. Text and audio files are stored in different directories as per female and male with their respective age group.

4.5.2 Praat:

This free software package developed, by Paul Boersma and David Weenink of the University of Amsterdam for speech in phonetics analysis that support all common platforms. PRAAT enabled services that we utilized in this work are: recording, synthesis, analysis along with articulatory synthesis.

4.5.3 Wave Surfer:

This is free and very simple voice editor software for acoustic phonetics analysis capable of interactive display for a given sample voice that includes pressure waveforms, spectral sections, pitch tracks and transcriptions etc. facilities available are to copy/paste, adjusting ZCR, normalize, echo, silence replacement, DC-removal etc.

4.5.4 Audacity:

This digital audio recording free source program is very helpful in speech analysis compatible with common operating systems developed by Dominic Mazzoni and Roger Dannenberg at Carnegie Mellon University in 2000. For post-processing task of audio files such as import/export of other audio format, mixing, normalization, trimming, and fading audacity can be used. Some of the main tasks that can be achieved are. Different kinds of digital effects, plug-ins support, Nyquist Spectrum Analysis (FFT based) etc are additional features audacity provides us.

4.5 Programming Software Tool-Matlab R2009a:

MATLAB developed by Math Works is a high-performance multi-paradigm numerical computing environment and fourth-generation programming language. In the late 1970s Cleve Moler, chairman of Department of computer science at the University of New Mexico began developing MATLAB. Although MATLAB stands for MATrix LABoratory, its suitability to handle other mathematical needs is well known and not just matrix manipulation. This interactive system's basic data element is an array that does not require dimensioning. MATLAB allows us to solve many technical computing problems, especially those which can have matrix and vector formulations, in a rapid pace it writes a program in a scalar non interactive language

such as C or FORTRAN. Beyond this MATLAB provides a family of add-on application-specific solutions called toolboxes. Toolboxes are comprehensive collections of MATLAB functions that extend the MATLAB environment to solve particular classes of problems like signal processing, control systems, neural networks, and many others. An additional package Simulink features graphical multi-domain simulation and Model-Based Design for dynamic and embedded system. The major advantages provided by MATLAB compared to other conventional computer languages for solving computational problems that make it suitable to use in our study are: Easy to use, Platform Independent, Availability of Predefined functions, Device Independent Plotting, Graphical User Interface, and MATLAB compiler etc. One of the most important and powerful toolboxes provided by MATLAB is *Signal Processing Toolbox*, which has been used in many occasions for different purposes to analyze the speech signals in the present study [70, 72].

4.6 Database creation, Acquisition and Organization:

Following figure 4.2 depicts the database tree structure used during the database creation for this research work.

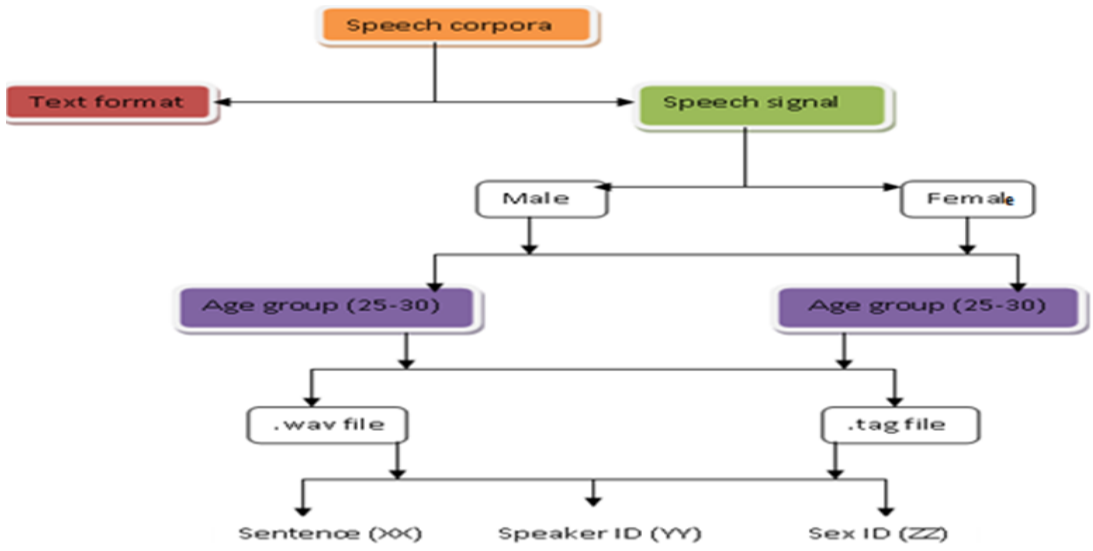

Figure 4.2: Database tree

4.7 **Annotation:** To keep track of the link between the speech signals and their linguistic representations on orthographic (and later on phonetic) level, the orthographic annotation of each recorded sentence is required. Precise annotation is important for corpus –based speech research. For the development of the BODO speech corpus, the entire task is divided into two phases: Phase I : The recording was transcribed by a skilled annotator and Phase II: The annotation was revised and corrected, if required, by another annotator:

- *RULES USED FOR ANNOTATION*: The recordings of speech from informants are done using the software Cool Edit Pro/ wavesurfer with 16-bit (PCM) resolution, sampling frequency 16 KHz. The segmenting, labeling will be done following Semi-Automatic process.

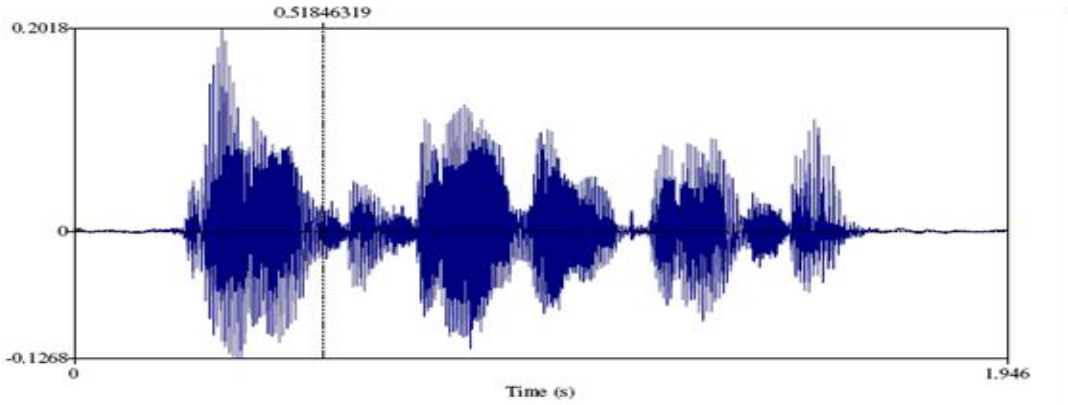

Figure 4.3: Labeling of speech signal.

During the annotation process, each sentence will be transcribed in the way it was really pronounced.

4.8 Policy for Tagging with some typical examples:

The annotations in prosodically annotated corpora typically follow widely accepted descriptive framewords for prosody. Usually, only the most prominent intonations

are annotated, rather than the intonation of every syllable. In the development of Bodo Speech Corpus, the following codes are used while tagging:

Tagged Description	Symbol
Phone	As per Phonetic Representation of Speech Labels
Syllable	^
Word	/
Phrase/Clause	,
Part of Speech	As per Tagging Identification for Parts of speech

The codes to be used for Part-of Speech Tagging of Bodo Speech Corpora are:

Symbols used	Description
AJ	Adjective
AT	Articles
AV	Adverb
CN	Coordinating conjunction
CJT	---------(THAT)
NC	Common Noun
NN1	Singular Common Noun
NN2	Plural Common Noun
NP	Proper Noun
PR	Pronoun
PRF	----------(OF)
PRP	Preposition
VBI	----------(be)
VM	Verb finite main
VI	Verb Intransitive
VT	Verb Transitive
VD	Past tense form of lexical verb
VG	----------(ing) form of lexical verb
VN	Past- particle form of lexical verb
IJ	Interjection

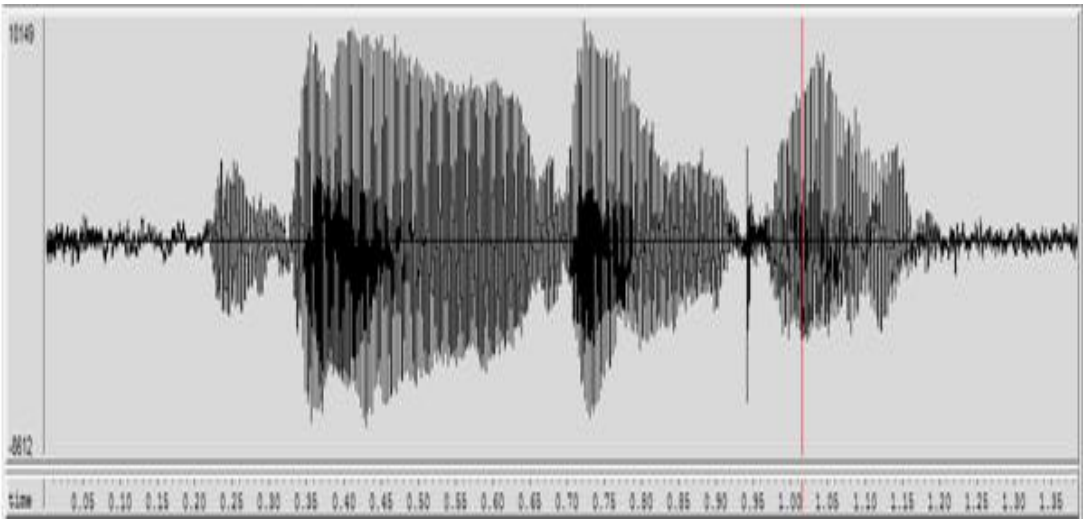

Figure 4.4: Bodo word बं-बं थां bong[AJ] bong[AJ] thAng[VT] -female.

Tagging Table of Bodo word बं-बं थां -female					
219	Si	#	#	#	#
332	B	#	#	#	#
503	O	#	#	#	#
660	Ng	^	/	AJ	#
332	B	#	#	#	#
503	O	#	#	#	#
660	Ng	^	/	AJ	#
680	Th	#	#	#	#
710	A	#	#	#	#
730	Ng	^	/	VT	#
760	Si	#	#	#	#

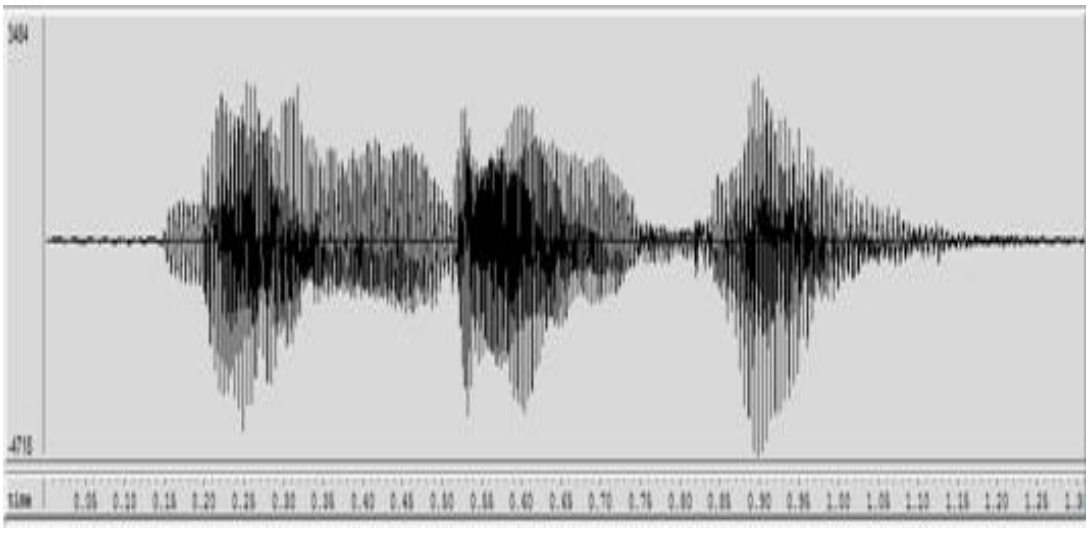

Figure 4.5: Bodo word बं-बं थां bong[AJ] bong[AJ] thAng[VT] -male.

Tagging Table of Bodo word बं-बं थां -male					
151	Si	#	#	#	#
205	B	#	#	#	#
336	O	#	#	#	#
509	Ng	^	/	AJ	#
547	B	#	#	#	#
619	O	#	#	#	#
759	Ng	^	/	AJ	#
877	Th	#	#	#	#
946	A	#	#	#	#
1132	Ng	^	/	VT	#
1310	Si	#	#	#	#

4.9 **Prosody in Bodo Language:** Prosody is defined as the suprasegmental information in speech, that is, information that can't be localized to a specific sound segment, or information that does not change the segmental identity of speech segments. For example, pattern of variation in fundamental frequency, duration, amplitude, or intensity, pauses and speaking rate have been found to carry information about such prosodic elements as lexical stress, phrase break, and declarative or interrogative sentence form. Similarly, hand gestures, eyebrow and face motions, can be considered prosody, because they carry information that modifies and can even reverse the meaning of the range lexical channel. Prosody used to carry emotion. For example, excitement is expressed by high pitch and fast speed, while sadness is expressed by low pitch and slow speed. Hot anger is characterized by over-articulation, fast, downward pitch movement, and overall elevated pitch. Similarly, cold anger shares many attributes with hot anger, but the pitch is set lower. Prosodic information is a source of information not available in Text-based system except punctuation.

4.10 **Intonation in Bodo Language:** Intonation is the process of intoning during utterance of speech. The system of significant levels and variations in pitch

sequences within an utterance takes place due to variation in Intonation. One such example is the type of pitch used at the end of a spoken sentence or phrase to ask a question is with a rising intonation.

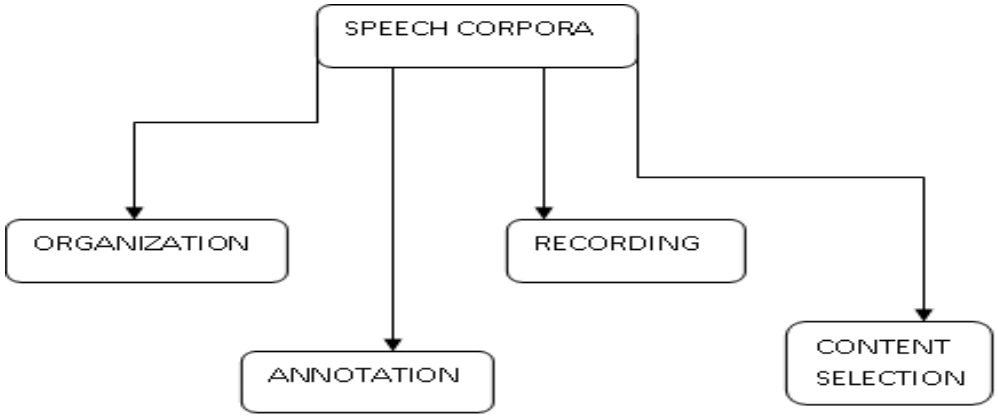

Figure 4.6: Database structure

4.15 Database Management Structure: All database contents posses' two interconnected components: Text component and speech component. The text components are stored as Unicode Text (16 bit) and the corresponding speech recordings are stored in PCM .wav format. The text and the speech information belong to the same filename with different file extensions. The wave file for individual word/sentence is stored in the form XXYYZZ.wav and the tag information file in the form XXYYZZ.tag file format, with XX as Sentence ID, YY as Speaker ID (age group) and ZZ as Sex ID (Male / Female)

The tag file contains: Time information, Symbols for phonemes (in Devnagari, Unicode conversion, IPA equivalent, Roman script), Syllables, Words, Phrase or Clause and sentences.

Chapter summary: In this chapter we have described the methodology followed to develop the proposed system SPERIA-B in detail. Along with that briefly database structure used is also explained.