Annotation of speech data

The term Annotation of Speech Data covers any indicative, descriptive or analytic notations applied to recorded raw speech data. These added notations include information of various kinds. Such recorded speech corpora must be tagged or annotated adequately for both linguistic as well as acoustic criteria.

4.1 Annotation:

Acoustic tagging consists of marking of (a) phoneme, (b) syllable, and (c) word boundaries along with the appropriate (d) Parts of Speech (POS) and/or (e) phrase/clause markers in the sound spectrum. Usually, there exists very little unanimity towards the definitions of phoneme-boundaries in any continuous speech. However, in the present task, the following conventions have been followed for marking phonemes:

- Consonant phonemes are defined from the beginning of occlusion to the end of Voice-Onset-Time (VOT) for plosives and affricates. Inclusion of VOT here means inclusion of the aperiodic transition, which, inter alia, is co-articulatory in nature; and, therefore, may vary for the same consonant in different context.
- For nasal consonants, trills, laterals and sibilants are defined as the durations of closure or constriction as the case may be.
- Vowel phonemes are defined as the total vocalic region inclusive of the vocalic transitions.
- For Vowel-to-Vowel (VV) transitions the segment boundary is placed at the middle of the transitory part.
- For diphthongs the segment includes the prominent target and the long transition.

For tagging sentences of the speech corpora, the sound spectrum should be tagged in terms of (a) phoneme, (b) syllable, (c) word, (d) phrase/clause and (e) Parts of Speech (Table 4.1).

Entire content of the corpora are stored in Unicode-encoded text with XML-directed front-end for access and search over the Internet. Speech signals of the corresponding word or sentences are stored in ".wav" file format under names decided as per the corpora management procedure. Each ".wav" file has an associated tag information file (with ".tag" extension), under the same name in text format. The phoneme markers to be placed on the sound spectrum are maintained as time offsets in the tag file.

Table-4.1 Tagging Policy of Annotation

Tagged description	Symbol
Phone	As per Phonetic Representation of Speech Labels (i.e. PASCII)
Syllable	"^"
Word	"~"
Phrase/Clause	","
Parts of Speech	As per Tagging Identification for Parts of Speech

The associated Tag File also contains special symbols corresponding to Phoneme Category, Syllable, Word, Clause and/or Parts of Speech as per Table 4. A completely Tagged Sentence is available for reference as Figure 4.1.

Table-4.2 Tagged File of Annotated Sentences: (ApnaAr ribolbHaArt.aA baAr korun)

Time Offset (in milli sec.)	Phoneme	Syllable	Word	Parts of Speech	Phrase/ clause
0	#	#	#	#	
166	sl	#	#	#	
274	aA	#	#	#	
387	p	^	#	#	
419	n	#	#	#	
509	aA	#	#	#	
531	r	^	~	PRP	
546	r	#	#	#	
656	i	^	#	#	
708	bH	#	#	#	
790	a	#	#	#	
858	l	^	#	#	
897	b	#	#	#	
999	aA	#	#	#	
1032	r	^	#	#	
1126	t.	#	#	#	
1179	aA	^	~	NNC	
1232	b	#	#	#	
1379	aA	#	#	#	
1427	r	^	~	NNC	
1522	k	#	#	#	
1602	o	~	#	#	
1623	r	#	#	#	
1753	u	#	#	#	
1924	n	^	~	VFM	
1998	#	#	#	#	

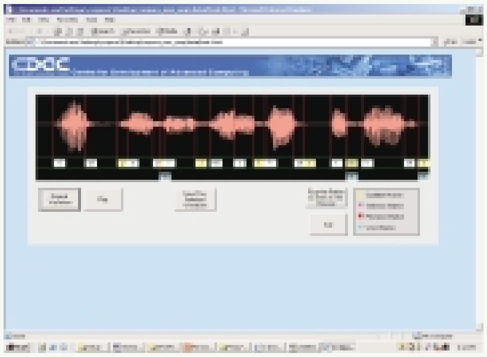
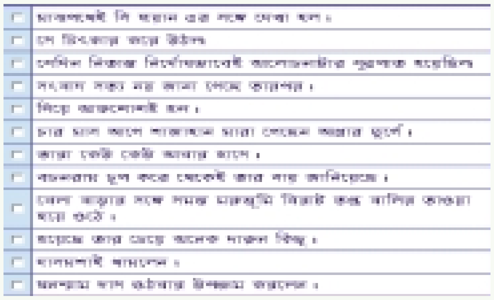


Figure 4.1: Tagged Wave Spectrum in wave viewer (aApnaAr ribolbHaArt.aA baAr korun)



- □ [illegible]
- □ [illegible]
- □ [illegible]
- □ [illegible]
- □ [illegible]
- □ [illegible]
- □ [illegible]
- □ [illegible]
- □ [illegible]
- □ [illegible]
- □ [illegible]
- □ [illegible]

4.2 Phonetic representation of Speech Labels:

Being a multi-lingual country, there is a need for standardization in symbolism representing various phonemes for such Speech Corpora management in India. In general, a highly optimized Grapheme Scripting Structure around "DECNAGARI" exists for written texts. This alone, or the International Phonetic Association (IPA) IPA set, could have served the standardization needs. Keeping in mind the National Need for a standard set of phonetic notations for all Indian Languages and also the ease of following such notations by an ordinary linguist, a scheme has been decided that uses the available character and symbol set of the Windows Platform.

In deciding the ASCII character set for such phonetic representation (named PASCII), both requirements of tagging, as well as typing into the computer of the phonetic scripts corresponding to the associated grapheme text, need be the guiding considerations:

The building of corpora for Indian Spoken Languages is a National Task. The IPA character representation of its standardized Tag Set is neither available fully on PC nor familiar to the Data Entry Operators, whereas the characters available with normal QWERTY keyboards are widely familiar to these Data Entry Operators. Furthermore, these are computer-programming-friendly and, therefore, processing them in PCs and/or in Databases become simpler.

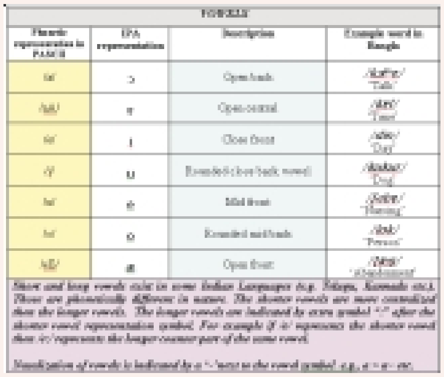
For ease of operation the additional attribute markers, like Tags indicating nasalization, centralization, retroflexion, etc. are taken only from the computer-processable Standard Character Set and placed adjacent to the respective phoneme.

The phonetic representation of some of common consonants of the Indian Languages require more than one character; for example, it is necessary to distinguish between /k/ and /kh/, between /t/ and /th/ along with the retroflexed versions of them. The necessity of using of two or more letters again may give rise to ambiguity during computer processing. It is, therefore, advisable to have identifiable characteristics for chosen notations of these phonetic tags. As frequency of occurrence of single letter consonants is generally high, it has been decided to use the small letters for the first letter and capital letters for the additional ones in the phonetic tags. For example, "k" stands for /k/ and "kH" stands for /kh, whereas "h" stands for only /h/.

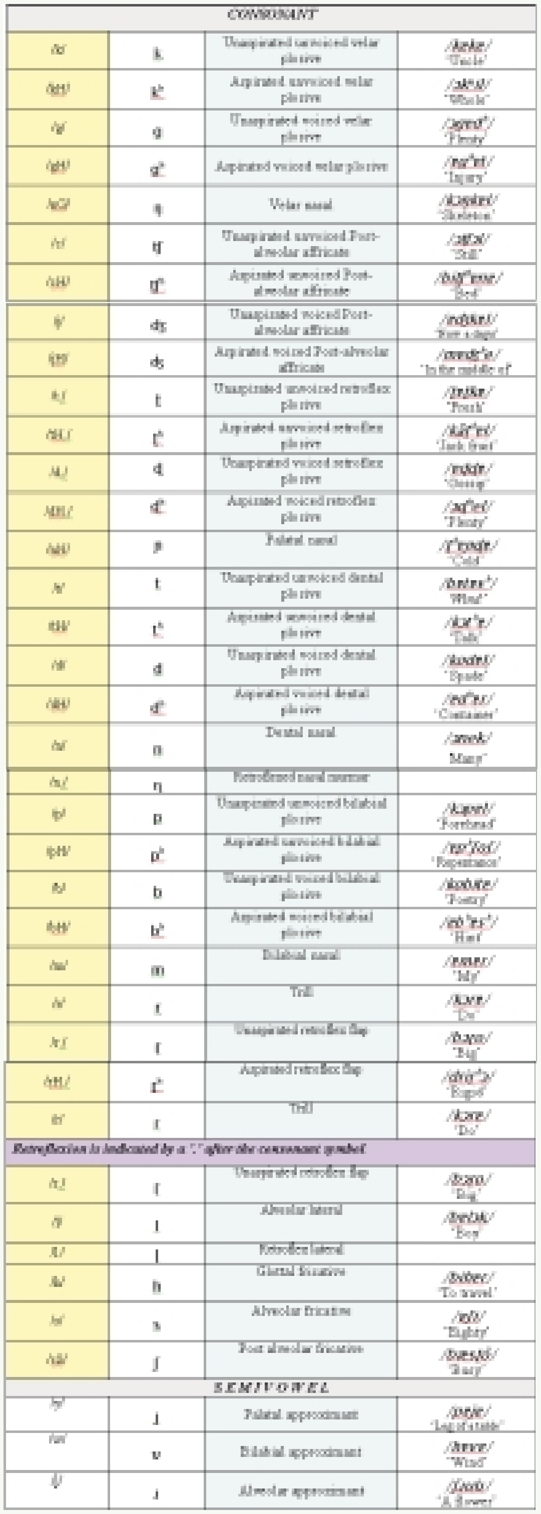
While some of the vowels in Indian spoken languages may have some small shifts in the position in the vowel diagram, in general, they may be represented by similar roman characters with some additional characters representing these special attributes as and when necessary. Thus, æ of the common IPA notations for Bengla becomes eE in the proposed notation; and similarly, kàk, khà   kH, etc.

The Table 4.3 describes the Phonetic representation of different phoneme for Indian languages in PASCII format. The table is based on the Bangla language and some other Indian languages. In addition special set of phonemes may need to be added for covering the all Indian Language

Table 4.3 Phonetic representations of speech labels



VOWELS			
Phonetic representation in PASCII	IPA representation	Description	Example word in Bangla
/a/	ɔ	Open back	/kaTHa/ 'Talk'
/aA/	ɐ	Open central	[illegible] 'Time'
/i/	i	Close front	/din/ 'Day'
/u/	u	Rounded close back vowel	/kukur/ 'Dog'
/e/	e	Mid front	[illegible] [illegible]
/o/	o	Rounded mid back	/lok/ 'Person'
/eE/	æ	Open front	/tEag/ 'Abandonment'
Short and long vowels exist in some Indian Languages e.g. Telugu, Kannada etc. These are phonetically different in nature. The shorter vowels are more centralized than the longer vowels. The longer vowels are indicated by extra symbol ":" after the shorter vowel representation symbol. For example if /e/ represents the shorter vowel then /e:/ represents the longer counter part of the same vowel. *Nasalization of vowels is indicated by a "~" next to the vowel symbol, e.g., a~ = ã etc.*			



CONSONANT			
/k/	k	Unaspirated unvoiced velar plosive	/kaka/ 'Uncle'
/kh/	kʰ	Aspirated unvoiced velar plosive	/akʰil/ 'Whole'
/g/	g	Unaspirated voiced velar plosive	[illegible] 'Thirsty'
/gh/	gʰ	Aspirated voiced velar plosive	[illegible] 'Injury'
/nG/	ŋ	Velar nasal	/kɔŋkal/ 'Skeleton'
/c/	ʧ	Unaspirated unvoiced Post-alveolar affricate	[illegible] 'Still'
/ch/	ʧʰ	Aspirated unvoiced Post-alveolar affricate	/biʧʰana/ 'Bed'
/j/	ʤ	Unaspirated voiced Post-alveolar affricate	[illegible] [illegible]
/jh/	ʤ	Aspirated voiced Post-alveolar affricate	/moʤʰe/ 'In the middle of'
/t./	ʈ	Unaspirated unvoiced retroflex plosive	/ʈatka/ 'Fresh'
/th./	ʈʰ	Aspirated unvoiced retroflex plosive	/kaʈʰal/ 'Jack fruit'
/d./	ɖ	Unaspirated voiced retroflex plosive	[illegible] 'Gossip'
/dh./	ɖʰ	Aspirated voiced retroflex plosive	[illegible] 'Thirsty'
/nj/	ɲ	Palatal nasal	[illegible] 'Cold'
/t/	t	Unaspirated unvoiced dental plosive	/batas/ 'Wind'
/th/	tʰ	Aspirated unvoiced dental plosive	/kɔtʰa/ 'Talk'
/d/	d	Unaspirated voiced dental plosive	/kodal/ 'Spade'
/dh/	dʰ	Aspirated voiced dental plosive	/adʰar/ 'Container'
/n/	n	Dental nasal	/anek/ 'Many'
/n./	ɳ	Retroflexed nasal murmur	
/p/	p	Unaspirated unvoiced bilabial plosive	/kapal/ 'Forehead'
/ph/	pʰ	Aspirated unvoiced bilabial plosive	[illegible] 'Repentance'
/b/	b	Unaspirated voiced bilabial plosive	/kobita/ 'Poetry'
/bh/	bʰ	Aspirated voiced bilabial plosive	[illegible] 'Hurt'
/m/	m	Bilabial nasal	/amar/ 'My'
/r/	r	Trill	/kɔra/ 'Do'
/r./	ɽ	Unaspirated retroflex flap	/bɔɽo/ 'Big'
/rh./	ɽʰ	Aspirated retroflex flap	[illegible] [illegible]
/r/	r	Trill	/kɔra/ 'Do'
Retroflection is indicated by a "." after the consonant symbol			
/r./	ɽ	Unaspirated retroflex flap	/bɔɽo/ 'Big'
/l/	l	Alveolar lateral	[illegible] 'Boy'
/l./	ɭ	Retroflex lateral	
/h/	ɦ	Glottal fricative	[illegible] 'To travel'
/s/	s	Alveolar fricative	/aʃi/ 'Eighty'
/sh/	ʃ	Post alveolar fricative	[illegible] [illegible]
SEMIVOWEL			
/y/	j	Palatal approximant	/paja/ 'Leg of a table'
/w/	ʋ	Bilabial approximant	/hawa/ 'Wind'
/J/	ɹ	Alveolar approximant	[illegible] 'A flower'

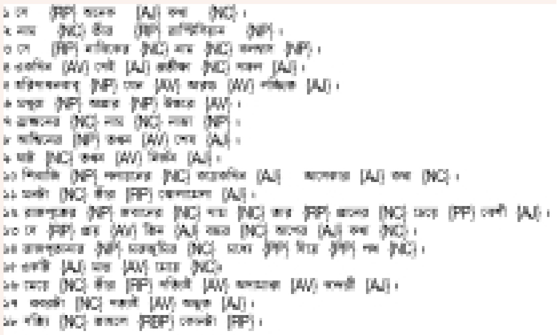
4.3 Parts of Speech Tagging

For prosodic analysis, all the sentences are tagged in terms of Parts of Speech (POS) with additional labeling on and above other tagging. These additional Parts of Speech tagging are done for carefully selected sentences and dialog by an experience linguist (like table 4.3). Table 4.4 represent the tag identifiers for Parts of Speech.

Table 4.4 Tagging Id for Parts of Speech

Sl no.	Description	Tagging ID
1	Common Noun	NC
2	Proper Noun	NP
3	Pronoun	PR
4	Adjective	AJ
5	Adverb	AV
6	Verb Finite Main	VM
7	Verb Auxiliary	VX
8	Verb Non Finite	
	Verbal Noun	VN
	Past Active Participle	VP
	Infinitive	VI
	Conditional Conjunctive	VC
9	Connective	CN
10	Post Position	PP
11	Interjection	IJ

Table 4.3



১ [illegible] [RP] [illegible] [AJ] [illegible] [NC] ।
২ [illegible] [NC] [illegible] [RP] [illegible] [NP] ।
৩ [illegible] [RP] [illegible] [NC] [illegible] [NC] [illegible] [NP] ।
৪ [illegible] [AV] [illegible] [AJ] [illegible] [NC] [illegible] [AJ] ।
৫ [illegible] [NP] [illegible] [AV] [illegible] [AV] [illegible] [AJ] ।
৬ [illegible] [NP] [illegible] [NP] [illegible] [AV] ।
৭ [illegible] [NC] [illegible] [NC] [illegible] [NP] ।
৮ [illegible] [NP] [illegible] [AV] [illegible] [AJ] ।
৯ [illegible] [NC] [illegible] [AV] [illegible] [AJ] ।
১০ [illegible] [NP] [illegible] [NC] [illegible] [AJ] [illegible] [AJ] [illegible] [NC] ।
১১ [illegible] [NC] [illegible] [RP] [illegible] [AJ] ।
১২ [illegible] [NP] [illegible] [NC] [illegible] [NC] [illegible] [RP] [illegible] [NC] [illegible] [PP] [illegible] [AJ] ।
১৩ [illegible] [RP] [illegible] [AV] [illegible] [AJ] [illegible] [NC] [illegible] [AJ] [illegible] [NC] ।
১৪ [illegible] [NP] [illegible] [NC] [illegible] [RP] [illegible] [RP] [illegible] [NC] ।
১৫ [illegible] [AJ] [illegible] [AV] [illegible] [NC]।
১৬ [illegible] [NC] [illegible] [RP] [illegible] [AV] [illegible] [AV] [illegible] [AJ] ।
১৭ [illegible] [NC] [illegible] [AV] [illegible] [AJ] ।
১৮ [illegible] [NC] [illegible] [RBP] [illegible] [RP] ।

Corpora Management

Corpora contents are expected to support four kinds of research investigation; namely: (a) speech research, (b) automatic speech recognition, (c) speaker recognition, and finally, (d) speech synthesis. Selected corpora contents of each of the above kinds need to have two interrelated components; namely: text and speech. Such text components are stored as Unicode text and the corresponding speech recordings are stored in Pulse Coded Modulation (PCM) in .wav format.

The tagged notations of the corresponding word or sentences are generated as per the corpora management structure. The text and speech information contain the same filename with different file extensions. While the speech recording files are kept in a separate directory as per corpora management structure, the corpora management (and the associated user interface) is designed using popular XML.

Common html navigation is used for locating particular information in the associated context. A search engine is also provided by which user can give a search for a particular word or a phoneme over the entire corpora. The search result display the corresponding html pointer for searched word; on right clicking over the word, one can get the corresponding speech file, containing that particular word. In case of phoneme search, it will provide the link ID as well as the position of the particular phoneme in the corresponding .wave file; and on clicking a particular button one can save the particular tagged phoneme signal.

There is a requirement of commonly known phonetic orthographic transcriptions of entire corpora content. One can also look at the phonetic transcription in IPA symbols of the orthographic representations by clicking appropriate button.

5.1 Corpora management structure:

Each of the four contents of the corpora can be selected separately for viewing. Each of these contents has two associated files; one for all the signals corresponding to the words and sentences in the .wave file format, and the other for the text file. The name of each file is generated using three fields 1) Field ID, 2) Speaker ID and 3) Word or sentence ID according to a convention depicted in Figure 5.1.

The wave file for individual sentence/word is stored in the from of XXYYZZ.wav and the tagged information file in the from of XXYYZZ.tag file format where XX = Field ID, YY = Speaker ID, ZZ = Word/Sentence ID. The tag file contains time information and symbols for the phonemes, syllables, words, clause/phrase and sentence as well as the Parts of Speech of each of the words.

The front-end is in an .html format and the text is stored in Unicode using XML format. Audio file information is stored in .wave format.

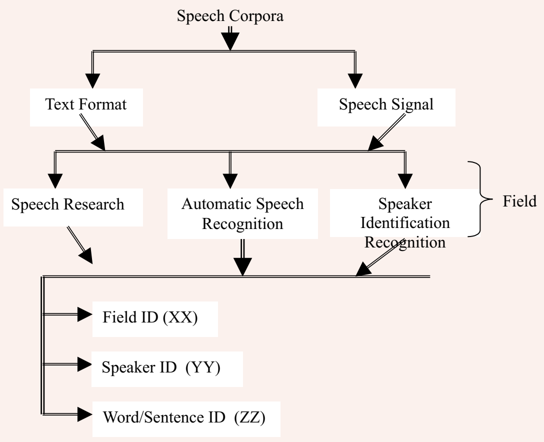


Figure-5.1 Corpora management tree

5.2 Corpora Management Methodology:

Corpora access must be easy and extremely user-friendly so that common linguistic researchers as well as corpora developers can intuitively view the desired contents and retrieve the whole or part of them without any special skill or in-depth computer knowledge. It means that suitable software needs to be provided; and over all management methodology must support such conveniences.

The corpora management should also contain the associated metadata information and it should present such metadata also in most user-friendly rendering.

To develop a user friendly front-end for corpora management, a navigation strategy has to be implemented. Figure 5.2 presents the adopted navigation strategy. The present corpus uses a simple html based hyper linked navigation for the user; where as the data is stored in .xml format. User can navigate to any intended location by simply clicking the appropriate link and reach the desired speech data through the common wave viewer. By clicking the appropriate link, one can down load, zoom, select specified portion; and also listen to sound corresponding to any of these. The wave viewer also provides for viewing of the speech wave along with its multi-level tagged information.

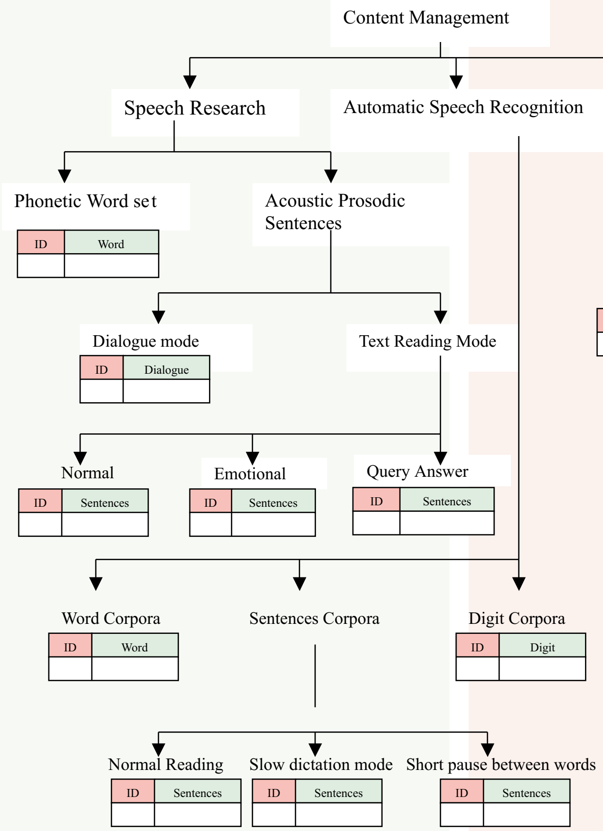


Figure 5.2 Corpora management diagram

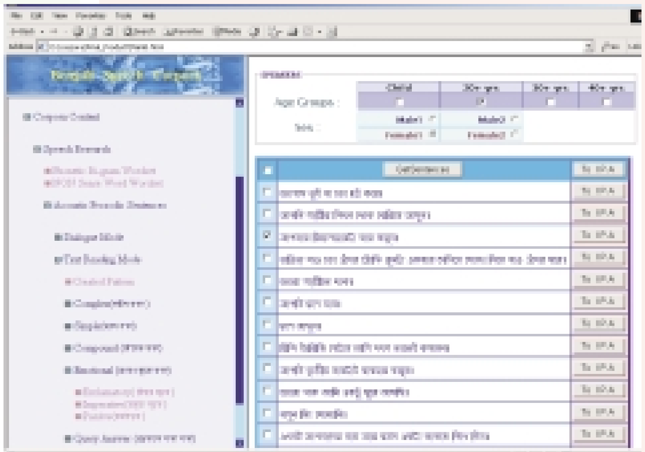
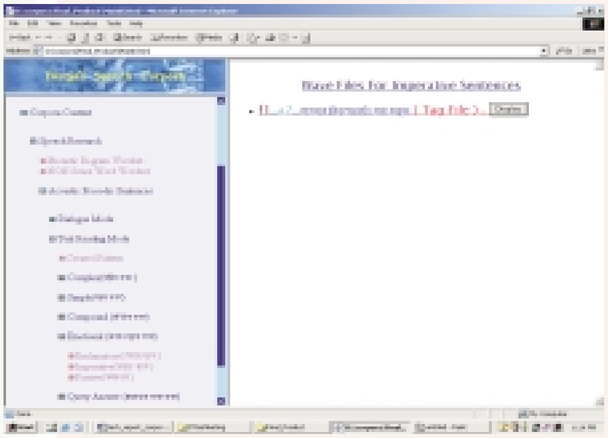


Figure 5.3 An Example of navigation in Bangla Speech Corpora



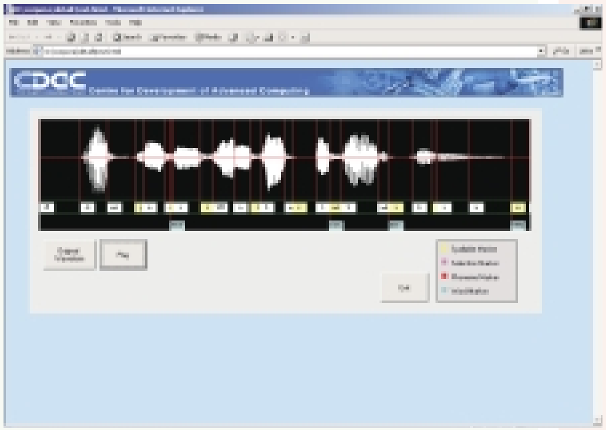
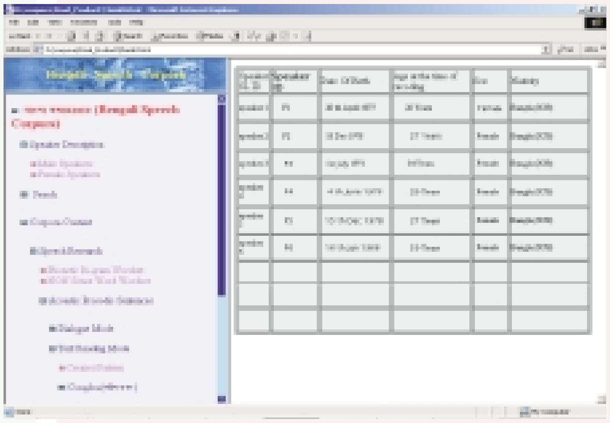


Figure 5.4 An Example of metadata display in corpora



Corpora Validation [1]

6.1 Corpora Validation Procedure:

Pre-validation of a small database of about 10 speakers shortly after the design specifications have been established and the recording platforms installed. The objective of this stage is to detect serious (design) errors before the actual recordings start. This stage also allows partners to build their database compilation software in an early stage of the project.

Validation of the complete database: The database is checked against the Speech corpora specifications and a validation report is edited.

Revalidation of a database: In case the validation report shows that corrections of a database are necessary or desirable, then (part of) the database can again be offered for validation, and a new report is written.

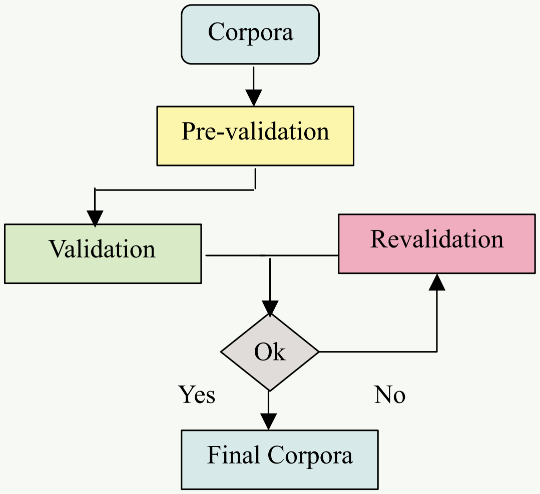


Figure 6.1. Corpora validation procedure

6.2 Checked List for Corpora Validation:

I. Documentation: It is checked if all relevant aspects of a Corpora (see II-VIII below) are properly described in terms of the three C's: Clarity, Completeness and Correctness.

II. Database format: It is checked if all relevant files (documentation, speech files, label files, lexicon) are present in the appropriate directory structure and with the correct format.

III. Design: The appropriateness of the recorded items for the purpose of the envisaged application(s) and the completeness of the recordings should be checked.

IV. Speech files: The acoustical quality of the speech files is measured in terms of (e.g.) (average) duration, clipping rate, SNR, mean sample value. Also auditory inspection of signal quality belongs to this category.

V. Label files: The label files should obey the correct format. Ideally, they can be automatically parsed without yielding erroneous information.

VI. Phonemic lexicon: The lexicon should contain appropriate phonemic (or allophonic) transcriptions of all words in the orthographic transcriptions of a SLR.

VII. Speaker & environment distributions: The recorded speakers should present a fair sample of the population of interest in terms of (typically) sex, age and dialectal background. Also the recording environments should be representative for the targeted applications.

VIII. Orthographic transcriptions: A (native) speaker of the language should check a sufficiently large sample of the orthographic transcriptions by comparing these to the speech in the signal files and the transcription protocol.

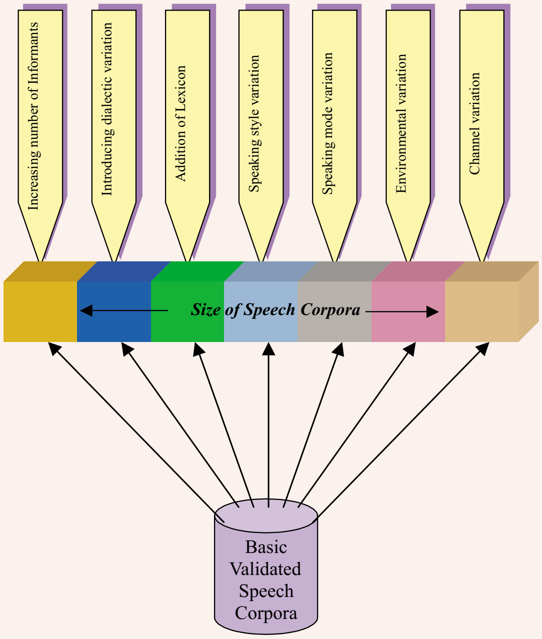
Conclusions

The aforesaid Speech Corpora is considered essential to support for the technology research and development for man-machine-man communication in Indian Languages. This Corpus may appear to be adequate only in this limited sense. Primary restriction is towards the manner of speaking (i.e. Text Reading Mode). The database for the study of supra-segmental characteristics may be inadequate for deriving rules for the purpose of generation of fully natural speech. It may be extremely difficult to mimic spontaneous human speech mechanically. Spontaneous interactions between speakers are not yet included. Furthermore, emotionally rich contents are deliberately avoided. There is a need for such corpora to contain normal spontaneous speech in future for both (a) broad based spoken language processing (SLP) as well as (b) linguistic studies on speech. Towards this end corpora may also contain, inter-alia, the following:

1. Normal spontaneous dialogue.
2. Natural query-answer on various topics of importance.
3. Inclusion of paragraphs on different emotions (data could be collected inter alia from cinema dialogues).
4. Inclusion of spontaneous speech (memorable events in the life of informants).

Possible extensions to the Speech Corpora:

Developed speech corpus examplifies the process of building speech corpora for Indian languages. For comprehensiveness, it may be advisable to increase the informants. The above corpus is developed for only one of the official dialect for selected languages. However, it may be necessary to build corpora covering the different popular dialects of the languages. There is also a need for special type of corpora to exemplify the speaking style of different speakers. In speech technology development (especially for Speech Recognition), the environmental noise and channel noise also adds important attributes; hence, there may be a need that corpora contain sufficient samples of that nature also. The content for acoustic prosodic study in the present corpus is only for one mode of talking. The corpus may need to be extended to include other modes (like anger, happiness, joy etc.)



Reference

[1] European Lang Resources Assoc. http://www.icp.grenet.fr/ELRA/

[2] Linguistic Data Consortium. http://www.ldc.upenn.edu/

[3] Shyamal Kr. Das Mandal, A.K. Datta, "Annotated Speech Corpora Creation for Bangla", Proceedings of SIMPLE-04, IIT Kargapur March 2004

[4] Shyamal Kr. Das Mandal, A.K. Datta, "Annotated Speech Corpora Creation: An Approach" Proceedings of FRSM-2004, Annamli University, January 2004.

[5] Shyamal Kr. Das Mandal, Indranil Sarkar, Arup Saha and Asoke Kumar Datta, "Creation of Speech Corpora for Speech Research and Speech Technology Development: An approach", ICSLT-O-COCOSDA-2004, New Delhi, 17-19 Nov.20

[6] Oregon Graduate Institute. http://cslu.cse.ogi.edu/

[7] Petrovska, D., et al. "POLYCOST: A Telephone-Speech Database for Speaker Recognition," RLA2C, Avignon, France, April 20-23, 1998, p. 211-214. http://www.speech.kth.se/~melin/papers/rla2c_ply.ps

[8] Campbell, J. "Testing with The YOHO CD-ROM Voice Verification Corpus," ICASSP. Detroit, May 1995, p. 341-344. http://www.biometrics.org/REPORTS/ICASSP95.html

[9] Reynolds, D. "The Effects of Handset Variability on Speaker Recognition Performance: Experiment on the Switchboard Corpus," ICASSP, May 1996, p. 113-116.

GLOSSARY

NLP	Natural Language Processing
MT	Machine Translation
SLP	Spoken Language Processing
Semiotics	In general, semantics (from the Greek semantikos, or "significant meaning," derived from sema, sign) is the study of meaning, in some sense of that term. Semantics is often opposed to syntax in which case the former pertains to what something means while the latter pertains to the formal structure/patterns in which something is expressed (e.g. written or spoken).
Partneme	Part of a phoneme
Nasal Nominal Pair	Difference between two words due to the Nasalization of a particular vowel.
XML	Extensible Markup Languages
VOT	Voice-Onset-Time
Diphthongs	When the two vowel cluster forms a single syllable the four vowel can only be the last component of the cluster. The last component has either the same height of articulation
IPA	International Phonetic Association