TTS Corpus specification		
1	Recording Instrument (Lab. Environment)	Dynamic Mic. With frequency response 80Hz-20 kHz(equvalant to Shure, Sennheiser,etc.) **Preamp.: 30Hz-15kHz** Sound Card: Creative Gold
2	Recording Environment	Speech studio (SNR >= -45 dB)
3	Recording Format	16bit PCM Mono, 22.05 kHz *
4	Informant Selection	Standard ITU-T (Annexure-1) , Age should be 25-35.
	Speech rate	Medium
	Emotion	**Neutral**
	Style	Read out
5	No. of Informant	2(1Male & 1Female)
6	Contents	**Sentences**→Cover all the di-phone, syllable and most probable tri-phone at least 2 occurrence probability. About 1000 phonetically balanced (PB) sentences. **Paragraph** (at least 5 sentences): 10-20 which more or less covers different prosody variation (Desirable: 3 repetitions of same data) **Story:** 2-3 stories of 4-5 paragraphs.
7	Annotation Hierarchy	**Acoustic** →Phone, Syllable, Word *Note : the defination of the Phone, Syllable, Word boundry in contineous* *Speech as given in annotation guidele (Annexure-II)* **Linguistic**→ POS (Functional), Phrase, Clause

* Comment by Alan W Black
Note : the above data based only for building the Read out mode TTS in any method for Stylistic synthesis the require data base will be different

Annexure-1

Informant Selection procedure

STAGE 1: Short-listing Potential Voice Talent from 10-15:

a. Basic Requirements for a Voice Talent:
    i. Native speaker of SCB
    ii. Educated at least to the first degree level
    iii. Age between 25-35 yrs.
    iv. Used to reading in a natural way without halting
    v. Should be familiar with English to be able to pronounce borrowed vocabulary correctly
    vi. Ability to speak for long sessions
    vii. Good voice quality (see b)
    viii. Good diction (see b)
    ix. Added bonus if the person is a professional radio artist because they are used to speaking from a script and projecting their voices.
b. Parameters for Subjective Evaluation of voice quality:
    i. No discernable speech defects like lisping, nasal voice, breathiness or very high pitched voice
    ii. Pleasant voice quality (pleasant to hear)
    iii. Clear articulation: good diction, no idiosyncratic pronunciation (like /s/ instead of /ʃ/, or dental instead of retroflex stops)
    iv. Accent: No pronounced regional accent
    v. Command of language: really familiar with the language and its sounds, no mispronunciation of words to be read
    vi. Intonation: natural intonation, neither over dramatized nor very monotonous, no "anglicized" or pronounced "read" intonation while reading.
c. Script for Recording :
    i. The script for recording should be a sub-set of the content to be recorded.
    ii. Approximately 5 minutes of continuous text (paragraph) and should include some borrowed English vocabulary
d. Recording:
    i. The recording set-up should be the same as that of the final recording
e. Subjective Evaluation by native speakers of the language: The voices recorded should be rated by at least 8 native speakers along a 1-5 scale (1 being disagree and 5 agree) for the following. Mean Opinion Score (MOS) should be calculated based on their rating for each voice.
    i. The voice quality of the speaker is very pleasant to hear.
    ii. The speaker has a good command over the language
    iii. The speaker has a dull or monotonous way of speaking
    iv. The speaker has very clear diction and pronounces each word clearly and accurately.
    v. The speaker has an over-expressive manner of speaking as if play-acting
    vi. The speaker has an anglicized accent

        vii.   The speaker has a marked regional accent
       viii.   The speaker has a nasalized voice
        ix.   The speaker has a breathy voice
         x.   The speaker lisps at places especially at stops
        xi.   The speaker does not pronounce English words properly

2. **STAGE 2**: Selecting the Voice Talent from the short list: The short-listed 5 voice talents should be put through the second stage of selection and evaluation. In the second stage, the speech is analyzed objectively (c) and 2 speakers with highest scores are then put through a subjective evaluation (d)

   a. Script for Recording:
      i. The script for recording should be a sub-set of the content to be recorded
      ii. Approximately 20 minutes of recording including a mix of read sentences of different intonation (declarative, interrogative, requests, exclamations, etc.), lists, and continuous text (paragraph)
      iii. 3-4 control sentences should be repeated at the start, middle and end of the recording.

   b. Recording: (same as stage 1)

   c. Objective Scoring measures: Measurement of voice quality is a slightly grey area and depends upon the purpose for which the analysis is to be done. In the case of recording voices for TTS, a good measurement is generally that related to the periodic versus aperiodic nature of the voice. For this purpose, measures for f0, jitter and shimmer may be used. Also, the duration and intensity across the recorded database needs to be consistent, hence these should also be measured. Fricatives are another problematic area where considerable overlap may occur, so, these should be also carefully analyzed.

      i. Jitter and Shimmer: Use the inbuilt jitter and shimmer algorithms in PRAAT. These measurements are made for 6 long vowels (two instances each of [a], [i], and [e] for each voice) at a constant pitch, that is, where there is negligible variation in pitch. The compared frames for each vowel pair should have the same number of pitch periods.
      ii. In breathy (aperiodic) voices, f0 dominates the spectrum and this is shown in the first harmonic (H1) of the spectrum dominating the second harmonic (H2). For a good TTS voice, H1-H2 (measured for non-high vowels, at steady state) should always be negative. **NB: These vowels should not be taken from words where they follow a voiced aspirated consonant**. (Another measurement for f0 dominance is the spectral tilt after 2000 Hz, a very steep tilt is indicative of a breathy voice. Though this might be too much to measure easily and can be used only as a secondary parameter)
      iii. Average Intensity measurements across different sentences should fall within a range and there should not be very sharp differences in intensity within the same sentence.
      iv. Duration: The duration of control sentences should be checked for drastic differences in overall duration as well as the duration of vowels occurring initially, medially and finally.
      v. Fricatives: the noise or aperiodic portion of all fricatives should be distinct and not overlap too much.

   d. Final objective scoring: A final score should be calculated for the objective parameters as follows:

i.   Jitter and Shimmer: the threshold for this is provided in PRAAT (you can use **Jitter (ppq5)** and **Shimmer (apq3)**). The lower the value of Jitter and Shimmer, the higher the rank.

ii. Rank the speaker according to their average H1-H2 values taken randomly for 20 tokens of non-high vowels. The negative values will be ranked higher than the positive values (Voices with positive values should be discarded)

iii. Rank the voices according to the range of intensity across sentences. Smaller range is rated higher than larger range.

iv. Rank the voices according to the difference in duration of control sentences as described in c-iv. Lower the difference, higher the rank

v.   Cluster the fricatives according to their place of articulation and average the frequency band in which the noise occurs. Check for the overlap between the different clusters. Rank the voices according to the overlap between clusters. Higher the overlap, lower the rank.

Subjective Scoring for voice quality: After the final objective scoring (d), evaluate the top 2 voices subjectively as in stage 1.

Annexure-II

Annotation Guideline

The every recorded sentence of the corpora is to be tagged in phone, syllable, word, in acoustic level and POS of the word phrase, clause boundary at linguistic level. Speech signal of the corresponding sentences is to be stored in ".wav" file format with a name decided by the corpora management procedure. Each ".wav" file has an associated tag information file, ".tag", in the same name of the sound file. The phoneme markers to be placed on the wave file are kept as time value [in ms] in the tag file.

Tag symbol for annotation

Phone	As per Phonetic representation of speech segment
Syllable	May be marked by '^'
Word	May be marked by '&'
POS	As per the deceided Annotation symbol
Phrase	May be marked by '/'
Clause	May be marked by '$'

Definition of phone boundary in acoustic signal

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 may be follow for marking phonemes.

a) Plosives and affricates

Phonemes are defined from the beginning of occlusion to the end of Voice-Onset-Time (VOT)[as Figure 1]
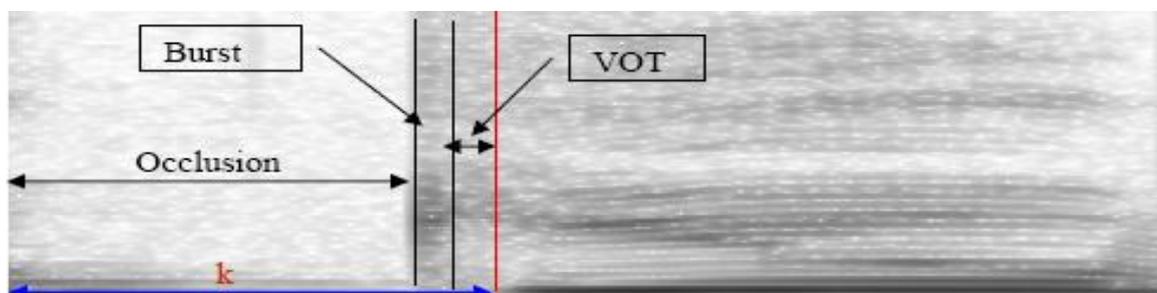

Figure 1. An example tagging of Plosives and Affricates consonant segment

b) Nassal murmur, Tril, laterals and sibilants

For nasal consonants, trills, laterals and sibilants are defined as the duration of closure or constriction [Figure 2]
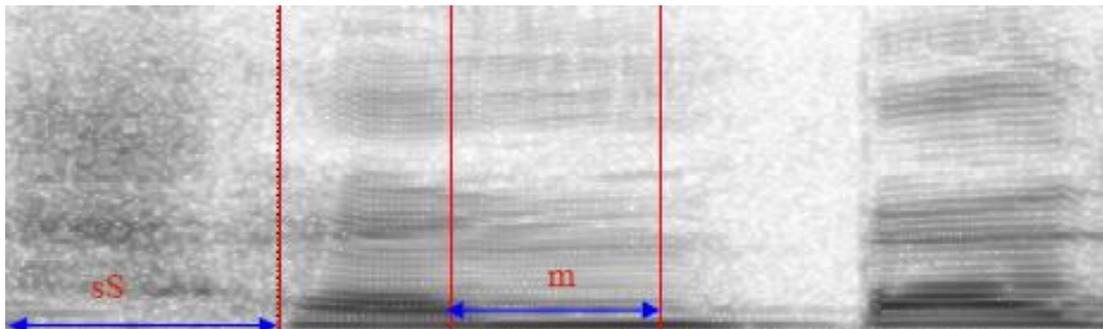

Figure 2. An example tagging of sibilants consonant segment

c) Vowel

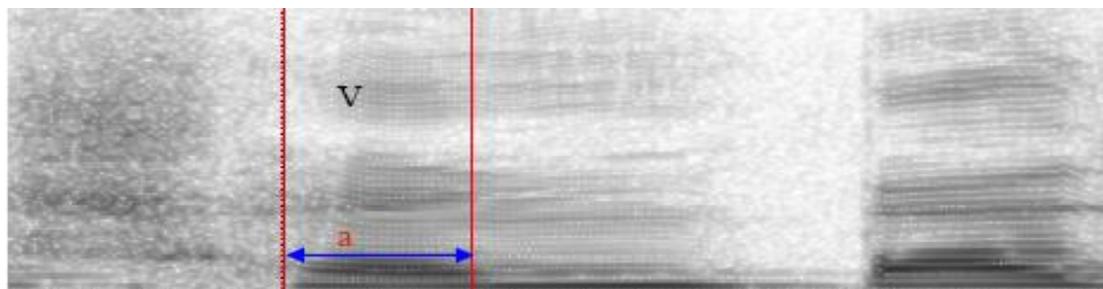Vowel (V) phonemes are defined as the total vocalic region inclusive of the vocalic transitions(Figure-3).

Figure 3. An example tagging of a Vowel segment

d) Vowel-Vowel Combination not forming dipthong

For Vowel-to-Vowel (VV) transitions if both the vowels are not part of the same syllable the segment boundary is placed at the middle of the transitory part (Figure-4).
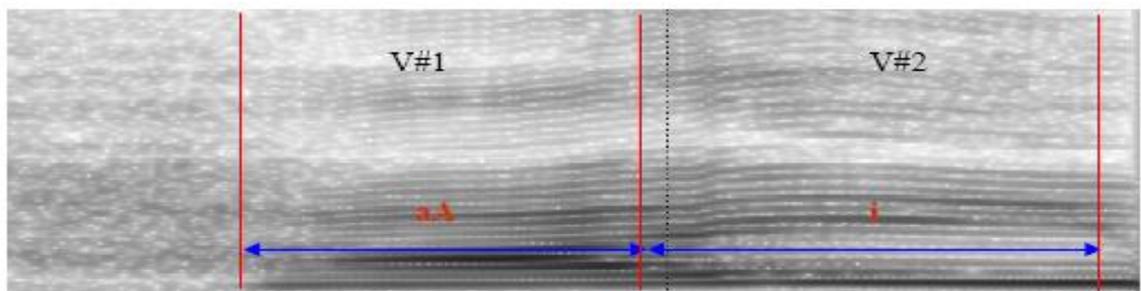

Figure 4. An example tagging of a Vowel-to-Vowel segment

e) Dipthong

For Vowel-to-Vowel (VV) transitions if both the vowels are part of the same syllable the segment boundary is placed at the end of the second vowel and the segment is marked as VV (Figure-5).
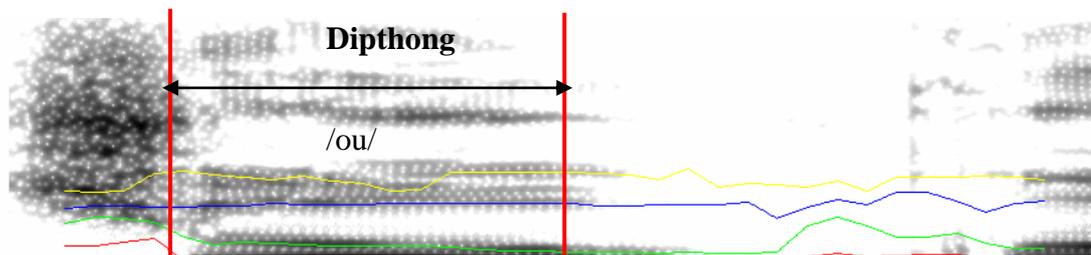

Figure-5: An example tagging of a Dipthong

f) Glides

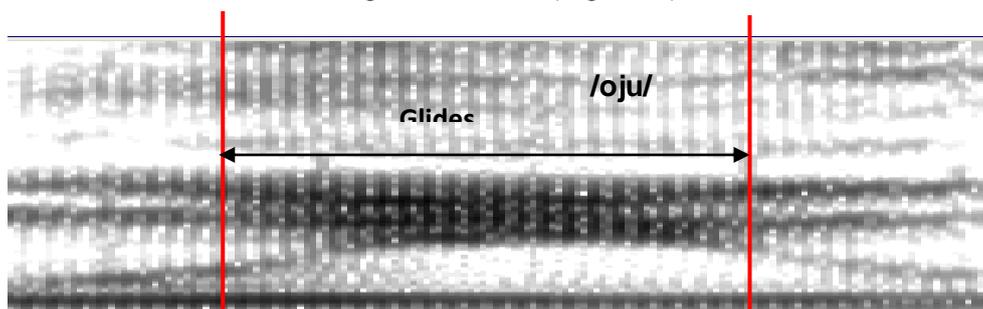Glides include transition to the contiguous vowels (Figure-6).

Figure-6: An example tagging of a Glides

The syllable and word boundray will be placed at the end of the corresponding phoneme boundary.

	Annotation Tools	Acoustic : G2P, Syllabification, Semi-automatic, Lingustic : POS tagger, Pharase and clause boundray marker.

Please review the above form for feedback and mail it to the following E-mail address
schandra@mit.gov.in