Text to Speech Testing Strategy

Version 2.1

07 July, 2014

Funded by

Technology Development for Indian Languages Programme
DeitY

Table of Contents

CHAPTER I: Testing Methodologies

1. INTRODUCTION

Text to Speech (TTS) is a system which takes text as input and synthesizes synthetic speech. The aim of testing and evaluation of TTS system is to judge the speech quality in terms of its similarity to the human voice and by its ability to be understood. Text to Speech system can be deployed on various platforms like:

a. Desktop Utility: One of the applications for desktop is TTS integrated with screen reader like NVDA (windows platform) and Orca (Linux platform).

b. Browser Plug-in: Plug-in extends the features of browser. TTS can be provided as browser plug-in which reads out the webpage to the user.

c. Handheld devices: On handheld devices TTS can be made available either in form of apps like Navigation, E-book reader or as its inbuilt feature.

2. PURPOSE

The purpose of this document is to propose a Testing & Evaluation strategy for the TTS system, which can be referred by any testing agency for determining the overall quality of a TTS system.

3. SCOPE

The scope of this document is to define TTS system testing strategy which mainly includes importance of test data, various testing techniques and subjective evaluation metrics. It also defines evaluator's selection criteria, and significance of appropriate training on how to assess "Speech quality of Text To Speech System".

Document scope is generic in nature taking into account the future development of TTS system. However, testing will be done as per the present limitation and specification of the system.

4. WHY TESTING IS IMPORTANT?

Software testing is an important phase in software development life cycle as it verifies and validates the system under test i.e. whether it works as expected and satisfies the stakeholders need.

With respect to TTS system also, testing & evaluation is significant; as it is important to test the system before deployment. For example: a screen reader application which is mainly designed for people with visual disability should fulfill their requirement of reading out the digital content. Testing & Evaluation of TTS system will be useful until the quality of synthetic speech becomes comparable to human voice; as most of the speech synthesis systems available today have distortions.

In order to assess the system output, appropriate quality assessment techniques should be adopted for determining the system performance in comparison to the benchmark level or with the quality of previous version or with similar kind of different product.

5. TESTING PROCESS

Testing is a process rather than a single activity. It starts with identifying testing scope of system, based on the defined scope of testing, test methodology to be followed is determined. One of the crucial factors in testing process is test data creation. Test data creation should follow the methods to be used in testing. If test data is wisely selected then testing will give meaningful results, thereby helping the developers/researchers in analyzing it.

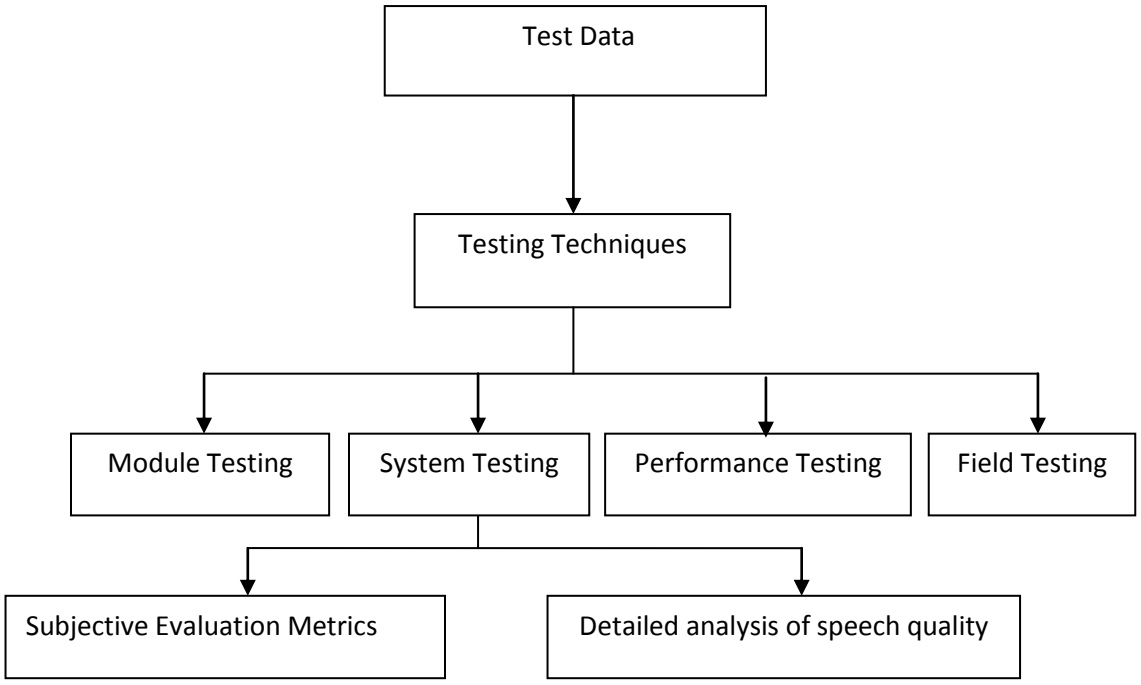

Fig. 1 Testing Process

5.1 Importance of Test Data

Test data plays a crucial role in testing; it can influence the whole testing process and can affect the test outcomes severely. Test data should be sufficient enough to cover all the functionalities of system under test.

A good test suite should be created against the system specifications, targeted domain and with respect to targeted application area. It should be designed in such a way that it covers all possible variations including numerals, abbreviations etc and language specific features like schwa deletion, geminates etc, which will help tester in verifying the system specifications, features, and in defining its limitations.

If test data is not focused & does not contain all such variations, then testing can take longer time to execute without generating any concrete results which may lead to meaningless testing.

For example:

In Generic TTS, data should be collected from different areas like history, science, tourism, health, news, stories, conversations, sports, weather forecasting, etc; it is to cover all the domains.

However, in Domain Specific TTS, test data collection should include targeted domain terminologies & contents. Through appropriate collection, test data coverage of domain can easily be checked. For example in the case of agricultural domain TTS, test data should cover crops name, various seeds, pesticides, fertilizers, mechanism of cultivation etc. For more details on Test data, refer Chapter – II.

5.2 Testing Techniques

Focus of testing should not merely be on finding the defects in the software but it should also consider its non functional aspects. Different testing techniques should be employed to test the system's quality and behavior, based on scope of the system under test. With respect to TTS system, Integrated System testing, Performance testing and Field testing are important and they are explained in detail below:

5.2.1 Module Testing

This type of testing requires knowledge of the module's internal structure and flow within the system under test. This is considered to be a useful testing technique for diagnosing the issues that may degrade the performance of specific module. In this testing each module is examined in isolation. Text-to-speech (TTS) systems usually contain following four modules.

One of them is **Text Preprocessing**, this is highly language dependent and it involves heuristic rules to handle different text preprocessing task. Major function of this module is expansion of textual features into its full words.

For example:

Textual expansions require contextual knowledge to correctly interpret and expand the text into its full word form; hence the efficiency of this module depends on the language dependent rules. For example: 1975 as Nineteen-seventy-five (if year) or one-thousand nine-hundred and seventy-five (if measure), 5/16 can be expanded as five-sixteenths (if fraction) or May sixteenth (if date).
For testing purpose, list of inputs with expected output should be made available and module errors can be drawn by comparing the module output with expected output [5].

The second module - **grapheme-to-phoneme conversion**, deals with the translation of written text into the corresponding phonemes set. The performance of grapheme to phoneme module can be expressed in terms of percentage of words correctly transformed into the valid phonemic string.

In the third module, computation of **prosodic markers** and their accumulation in phonemic string is done.

The last module deals with **production of the speech waveforms**.
On the whole, speech quality depends on the performance and output quality of individual module. Evaluation of each module is important as it help researchers or developers in determining the reason of shortness [13].

5.2.2 System Testing

In System Testing, a complete integrated system is considered for testing irrespective of internal working and code. Different functionalities of TTS system can be evaluated by examining overall output speech. Evaluation of linguistic aspects & overall speech quality evaluation are the two major objectives which can be fulfilled through integrated system testing.

Text preprocessing, segment handling and prosody are the broad features which are covered in Linguistic aspects evaluation. These features can be tested by giving input sentence e.g. declarative sentence to the system and observing the stress attribute in the overall speech output, by this prosody module can be evaluated.

From the end users perspective, evaluation of overall speech quality is utmost important as they are usually interested in the performance of a system as a whole that accepts text as input and generates corresponding speech. One of the known ways of doing this is through opinion testing. In opinion testing group of listeners are asked to rate the speech quality of a TTS system on various attributes like naturalness.

5.2.3 Performance Testing

Performance testing is an assessment of reliability of what is being tested. In this, performance specifications are verified and performance in terms of responsiveness and stability is also determined.

Responsiveness is considered as an important factor in determining the usefulness of system. It is the time taken by the system to respond the request. In TTS system, it is the time taken by the software in producing the output speech. In real time scenario consideration of response time is vital; in many cases poor response time can affect the functionality of application. If response time of a TTS navigation app which is designed to assist the directions is below the expected level, purpose of using such app becomes meaningless.

5.2.4 Field Testing

This test evaluates the performance of the system within the context of specific users and environment in which the system will be deployed and used. Field testing can be useful in identifying a wide range of interaction problems such as problems with software being incompatible with other software on the targeted systems, inadequate training to the intended users, unavailability of user manual etc.

Whereas laboratory tests are generally performed under the controlled conditions i.e. with low noise level. The primary difference between field and laboratory test is the environment under which the testing is performed. Laboratory test can be performed at different stages of software development life cycle, or it can be performed to verify the system's performance against its previous version. But sometimes it is not possible to predict the exact behavior of TTS system on the basis of laboratory test only; in such conditions it is essential to undergo field test. Though optional, but it is beneficial to perform field test whenever significant changes are introduced or when a new system is to be deployed [5].

With respect to TTS system also, field test is very meaningful e.g. TTS as screen reader application specially designed for people with disability should be tested by the intended user class; as success of such applications can only be judged in real environment. Field test should be performed only after successful completion of internal acceptance test in laboratory.

Ideally this test should be conducted in actual user space, but it can also be performed by simulating the user environment in the test lab and by asking the actual users to listen and give their acceptance for the system under test in terms of Yes or No [13].

5.3 Subjective Evaluation Metrics

5.3.1 Naturalness Test

i. MOS for naturalness

The Mean Opinion Score (MOS) is the simplest method to evaluate the quality of a speech synthesis system. MOS gives a numerical indication of the quality of the synthesized speech. In this method, focus of evaluator should be on naturalness of synthetic speech. By Naturalness it is meant that the sound is indistinguishable from human speech and it is close to human voice [2].

Method:

Evaluator will have to understand the naturalness aspect of synthesized speech after listening the voice played. MOS is the arithmetic mean of the scores given by all the evaluators.

MOS Scale

MOS	QUALITY
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Formula for calculation of MOS:

$$\mathbf{MOS} = \frac{\sum_{j=1}^{M} \left(\frac{\sum_{i=1}^{N} S_{ij}}{N} \right)}{M}$$

S_i = Score of i th evaluator

N = # of Evaluators

M = # of Sentences

j = Sentence index

Test Data for MOS:

Sentence level or short paragraph data that should cover all the possible variations mainly: numerals, abbreviations & acronyms, symbols, English words written in Latin and in Indian scripts and different types of sentences like declarative, negative, and exclamatory. Sentences should be selected from different areas namely news, stories, sports etc.

For example:

1. विवाह की तिथि १५ नवम्बर २०१३ को रखी गई है |
2. डॉ रस्तोगी ने कहा की वो अपना एटीएम भूल गए थे ।

ii. Degraded Mean Opinion Score (DMOS)

The quality of synthesized speech depends on the quality of the recorded voice. DMOS test is carried out to check the naturalness aspect of the speech by comparing the natural and synthetic voice.

Method:

In DMOS method, evaluators need to listen to synthetic as well as natural voice in random order without having prior information about the type of voice i.e. natural or synthetic, this is to avoid biased scoring. Motive of this method is to judge the voice in terms of naturalness. Average of scores given to natural sentences and synthesized sentences separately by each evaluator, will be calculated [4].

For each evaluator,

Normalized score of synthesized to natural = (synthetic voice score/natural voice score) * 5

System's DMOS score will be the average of normalized score of each evaluator.

Scale of Naturalness

5	System sounds like Human
4	Robotic sound but reading correctly
3	Reading sentences with less broken words in robotic manner
2	Almost every word broken
1	Extremely intolerable

Test Data for DMOS:

Quality of system generated synthetic voice largely depends on natural voice and its recording conditions. Domain, coverage and variations in synthetic voice should match with the natural voice used. If the content used while recording natural voice is of agricultural domain then the test data used for synthetic voice should be from similar domain and same complexity. This is to get the fair evaluation.

5.3.2 Intelligibility Test

Intelligibility is one of the important factors affecting speech quality. We can calculate intelligibility either by MOS and WER.

i. MOS for Intelligibility

The Intelligibility refers to the accuracy with which each word is pronounced so that normal listener can understand the spoken word or phrase. In this method, focus of evaluator should be on intelligibility of synthetic speech.

Above mentioned formula for MOS calculation will be used here also.

ii. Word Error Rate

Method & Test Data for WER:

For such test, sentences which are semantically unpredictable (SUS) but are constructed in such a way that they are grammatically syntactical should be used. Word length of SUS sentences should not exceed by 7; otherwise chances are that people might forget them. SUS is used to evaluate the intelligibility because it becomes difficult for listeners to predict the unheard information. After listening a sentence, evaluator will have to write whatever they heard, even if they don't understand the meaning. While calculating WER, typographical mistakes should be avoided. [4]

Memory Test should be conducted to check whether the evaluator is able to memorize the sentence played or not, since people with low memory can rate test poorly and to avoid this, evaluation should be performed by those who qualify the memory test.

For each person,

$$\mathbf{WER} = \frac{\mathbf{(S+D+I)}}{\mathbf{N}}$$

Where,
S is the number of substitutions,
D is the number of deletions,
I is the number of insertions,
N is the number of words in the Sentence

For overall SUS scoring, percentage average of WER will be calculated [9].

Example of SUS: रखे हिमालय कंधे पर, चली सूर्य।

5.3.3 Accuracy Test

For accuracy calculation proper selection of test data is crucial. All such data whose expected output is well defined can be considered for accuracy test. Categorization of test data is as follows:

a) Number Handling
   i. Digits (Phone number +91-9999999999)
   ii. Fractions (2/3)
   iii. Numbers (1004)
   iv. Numerals (1st, XII)

b) All Date formats (01/12/14, 9 जनवरी, 2014)

c) Foreign words (Latin script)
d) English words transliterated in Indian script (इंस्टिट्यूट)
e) Acronyms (etc., prof., डॉ, Rs.)
f) Abbreviations (SBI, DRDO, पीडब्लूडी)
g) Names
h) Addresses
i) Homographs
j) Punctuations and Brackets (, ; “ ” – [],(),{})
k) Special Symbols ($, @, %)

Punctuation handling is application dependent e.g. if the application is designed for people with disability then the system must readout all the punctuation marks.
For e.g. (True) it should read as →’ round bracket open’ ‘ True’ ‘round bracket closed’; applications designed for people other than PWD appropriate pauses for punctuations should be checked[16].

5.3.4 Comprehensibility Test

The word "Comprehension" means up to what extent the message received is understood. The performance of TTS system can also be evaluated through comprehensibility test. Merely by identifying each word it is difficult to get the deeper information about the context, because in intelligibility test evaluator will emphasize only on recognition of each word without concentrating on the meaning of the sentence. Comprehensibility test should be carried out when the system achieves the intelligibility up to acceptable level otherwise it is meaningless to carry out comprehensibility test for unintelligent system; as intelligibility has a strong impact on comprehension [10].

Method:

In Comprehensibility test, evaluator will be asked to listen to a paragraph/story (100 – 150 words) and based on that a series of questions will be posed. Questions should be framed in such a way that whether the evaluator has understood the paragraph heard or not can be observed.

A two point scale (0, 1) is recommended.

Scale	Remark
0	Incorrect response
1	Correct response

There are two kinds of questions that can be prepared for comprehension test.

1. Open ended questions: the evaluator will have to write down the answer.
2. Closed ended questions: the evaluator will have to choose an appropriate answer from multiple options, generally 2 or 4 options are provided.

Questions can be factual and inferential.
For comprehensibility test, data should be wisely created/chosen as there may be chance that evaluator has prior knowledge of the paragraph/story played and they may answer it based on their prior knowledge. Data for such test can be picked from stories of one language e.g.: Punjabi & translate it in required language e.g.: Hindi, by this method chances of prior knowledge of such data reduces.

5.3.5 Additional Aspects of Speech Quality Assessment

There are some aspects which are also important apart from naturalness and intelligibility of the synthesized speech they are mainly pronunciation, listening efforts and speaking rate. For observing all the aspects, evaluator will have to give their opinion on each aspect separately.
Scale for each parameter will be provided and listeners will be asked to choose single option from each scale [3].

In "Pronunciation Scale" concentration should be on finding the anomalies in synthetic speech pronunciation.

Scale of Pronunciation of words

☐	Sound of the word effortlessly understood
☐	Pronunciation of the word is more influence to native speakers
☐	Basic word pronunciation of the intended language
☐	Pronunciation by breaking words, overlapping of words, skipping words
☐	Not able to understand the word pronounced

In case of “Listening Effort”, efforts required to understand synthesized speech has to be observed.

Scale of Listening effort

☐	Complete relaxation possible; no effort required
☐	Attention necessary ; no appreciable effort required
☐	Moderate effort required
☐	Effort required
☐	No meaning understood with any feasible effort

For scaling “Speaking Rate”, listener will have to monitor the rate with which the synthesized speech is played.

Scale of Speaking Rate

+2	Much faster than preferred
+1	Faster than preferred
0	Preferred
-1	Slower than preferred
-2	Much slower than preferred

5.4 Speech Quality Evaluation Procedure

The first step of speech quality evaluation procedure is to evaluate the overall speech quality by various subjective evaluation metrics such as naturalness test, intelligibility test and comprehensibility test. Since these test are subjective in nature it should be evaluated by the real user of the system. Human perspective varies from one to another so large number of evaluators should be involved in order to get fair evaluation. This type of evaluation methods are time consuming and expensive.

Next step is to select valid evaluator's on the basis of evaluator selection criteria. Major step that can affect the evaluation result is the evaluation training and the effectiveness of the training depends upon how well evaluator's are trained.

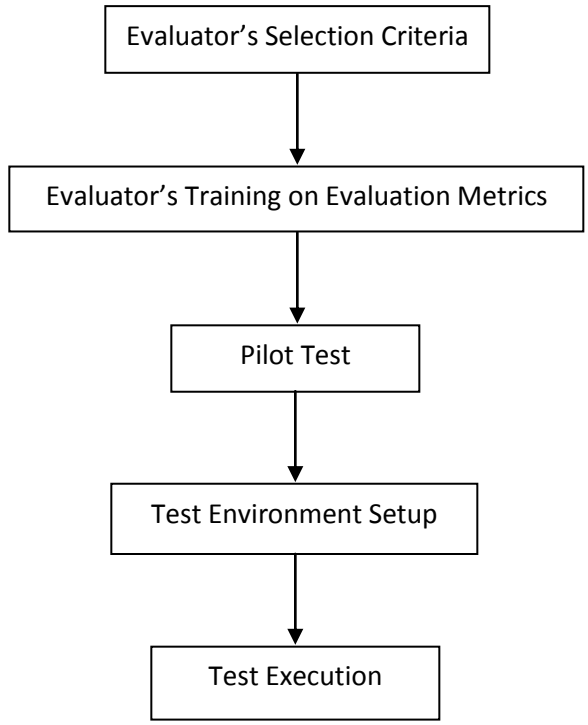

Fig.2 Evaluation Procedure

5.4.1 Evaluator's Selection Criteria

Criteria for evaluator's selection are as follows:

- They should be a native speaker of the assigned language for evaluation.
- Minimum education qualification should be high school.
- They should have basic mobile & computer knowledge.
- They should not have any medical history of hearing loss or impairment etc. To confirm this audiometric test should be carried out.
- They should not have participated in any listening test from previous 6 months; if they are exposed to the system recently then listeners will tend to predict the behavior pattern of synthetic speech without giving required efforts [4].
- Their age should not be less than 18 years.
- If the application is designed for a specific class of users then the evaluator's selection criteria should also meet intended user's specifications. For example: A story reading app should also be tested by the children of intended age group.
- Pool of evaluators should have people from various age group, different regions and should have equal percentage of both the gender.

5.4.2 Evaluator's Training Guidelines

5.4.2.1 Evaluator Training Overview

The aim of the evaluator's training session is to make them familiar with the evaluation metrics by giving overview of each metrics. Training session should be conducted language wise, in which examples for each evaluation metrics should be explained. Response on these metrics will be taken in following two ways from the evaluator:

I. **Grading Procedure**

In this procedure, evaluator will have to listen a message and based on the voice heard they will have to rate the system from the provided scale; that should reflect their opinion about the quality of the message played. Listener's focus should not be on sentence structure or grammar, just how it sounds. For better understanding minimum 2 examples for each evaluation method should be played.

II. Transcription of voice played

In this procedure, evaluator will be asked to listen message and then write what they heard, rather than understanding the voice played. In case, listener cannot understand any of the words, they should write "X" only.
For inputting the text, language keyboard may be provided and a keyboard hands on session can be conducted.

5.4.2.2 Pilot Test

Before conducting actual evaluation, a pilot test with the evaluators should be carried out beforehand. In this test the evaluator should have hands-on session of the evaluation system; this is to clear all major doubts regarding the evaluation process which can affect the evaluation results. Duration for pilot test should be kept short but sufficient enough to cover all examples.

5.4.3 Test Environment Setup & Test Execution

Testing environment is a setup of software and hardware on which the testing team performs the intended test. In case of TTS system the ambient noise level in the test laboratory should be kept as low as possible; this value should be close to the noise level in the hospital/libraries. Listeners should be provided with good quality headphones with their volume set to an adequate loudness [12], because the experimental conditions themselves affect the evaluation test results.

Evaluations are highly susceptible to the size of evaluator's pool, so it is important to use large number of evaluators to get more appropriate results. Evaluation test can be performed either manually or through evaluation assistance tool.

5.5 Detailed analysis of speech quality

There are some aspects which should be evaluated by linguists, like segments and prosody because evaluation of these aspects requires good observation skill in terms of language. To implement these tests one should have detailed language knowledge, so few linguists can carry out these tests.

5.5.1 Segmental Evaluation

With segmental evaluation methods single segments are recognized separately for intelligibility.

i. MRT : Modified Rhyme Test

From the name itself it is understood that it's a rhyming test in which intelligibility of word is checked. This test is to judge whether the evaluators can differentiate the rhyming words or not. At a time first letter of the word or last letter of the word is changed(can be substituted with multiple letters) but not both at the same time.

Method:
Audio of a word will be played to the evaluator and after listening the same, 4 options will be provided to them and they will have to choose the right option.

For e.g.: Audio of अगम will be played.

options can be a) अगम b) बेगम c) सुगम d) निगम

To check the alertness of evaluator slight modification can be done in the test. In the options, instead of keeping the correct word, "none of these" option can be included.

For e.g.: Audio of कलश will be played.

options can be a) कलमा b) कलकी c) कलरव d) none of these

ii. Difficult words: Word containing one or more conjunct which are difficult to pronounce. Difficulty level of such words increases with character length. Result of this test may be recorded in a tabular form, having Input word, correctly recognized syllables.

For example: Input Word: इंस्टीट्यूट

Syllables in word: इं + स्टी + ट्यू + ट

IPA representation: /ĩsʈiʈəjuʈ/

5.5.2 Language specific features evaluation

Language specific features like schwa deletion in Hindi, Marathi, compound words and Geminates in Tamil, Punjabi, Telugu.

i. **Schwa Handling:** Consonants are associated with inherent 'ə' (schwa), and it is not explicitly written in Indian languages. schwa is sometimes pronounced and sometimes not.For eg: सरल(sərəl) सरला(sərlɑː)

Proper handling of schwa is important. In some cases, depending on context same character-sequence can be pronounced differently. And if the system fails to handle schwa, it will be difficult for the listener to understand the intended meaning of the word. For eg: 'रक' is pronounced differently in हरकत (hə**r.k**ət,) and सरकना (sə**rək**.nə).

For Dravidian languages, schwa is not deleted while reading named entities. For eg: Ramanarayanan (rɑːmɑːnɑːrɑːjɑːnɑːnə).

ii. **Geminates:** it is also called as consonant elongation, means spoken consonant is pronounced for longer duration. In Punjabi language diacritic called an 'addak' which is written above the word and indicates that the following consonant is geminate.

For e.g: ਦਸ - Ten; ਦੱਸ -'Tell' (verb)

ਪਤਾ -'Aware of something' ; ਪੱਤਾ -'Leaf'

iii. **Compound Words:** Two or more words are joined together to create a word, meaning of compounded word may vary from isolated words. In Dravidian languages compound words are frequently used, while testing Dravidian languages such words should be collected.

For eg: **எழுதுகோல்** (Tamil)

iv. **Nasalization of word:** Devanagari script has nasalized vowels apart from ऋ and consonants (ङ ञ ण न म) .words are nasalized by using chandrabindu, anuswar, conjunct aksharas. Focus should be on proper identification of nasalized sound.

Nasalized vowel: हँसना, घड़ियाँ, सिंचाई, नहीं

Nasalized consonant: कंबल, जङ्गल, ठण्डा, गंदा

5.5.3 Prosody Evaluation

In TTS system one of the major tasks is to improve the naturalness of synthetic speech, which is mainly dependent on the quality of prosody module. Prosody affects naturalness & comprehensibility rather than intelligibility [14].

This module take cares of various properties in speech utterance namely intonation (word stress and pitch control), speech pauses (word & sentence boundaries) and emotions. These properties cannot be determined merely by identifying consonant and vowel phoneme set. Prosody carries some meaning of its own especially because of intonation, as it allows the system to present a sentence as a declarative statement or a question, or to express emotions to the listener.

So, there is a need to evaluate prosody module for determining the naturalness of speech output. For meaningful evaluation, textual test data should cover different types of sentences and sentences with different punctuation marks for examining speech pauses.

5.6 Comparative Testing

Comparative testing can be done by comparing the performance of various speech synthesizers. Same test suite should be used for generating output of different speech synthesizers. Synthesizers can be of different: vendors, versions of same product and development approaches of synthesizer. Ideally, following points should be considered while comparing:

1. To compare different version of same product, synthesized speech should be of same person's voice.

2. For product of different vendors, synthesized speech should be of same gender. Voice pleasantness is also an important factor which should also be checked.

Method:

Output of speech synthesizers should be presented sequentially to the listener and output files should be generated from same input text.

Repeated presentation of same text may produce a learning effect in listeners which may result in higher intelligibility with more positive attitude towards later systems. Higher intelligibility may result in biased rating; to avoid this, listeners should not have any idea about source synthesizer of voice played. It can be achieved by randomizing the sequence of various synthesizers' output. For example:

Input Sentence	**Output voice of different speech synthesizer(voice)**			
	Position_1	Position_2	Position_3	Position_4
Sen1 (मेरी उससे मुलाकात नहीं हुई है।)	Voice1_Syn_A	Voice1_Syn_B	Voice1_Syn_C	Voice1_Syn_D
Sen2 (क्या तुम स्कूल जाते हो ?)	Voice2_Syn_D	Voice2_Syn_A	Voice2_Syn_B	Voice2_Syn_C
Sen3 (सुबह उठ कर व्यायाम किया करो।)	Voice3_Syn_C	Voice3_Syn_D	Voice3_Syn_A	Voice3_Syn_B
Sen4 (शायद पिताजी आज आ जाये।)	Voice4_Syn_B	Voice4_Syn_C	Voice4_Syn_D	Voice4_Syn_A

6. Selection of Testing Methodologies

In the above section all the testing methodologies has been explained but while evaluating TTS system selection of appropriate methodology is very important. Selection of evaluation method depends on system maturity and its features.

Testing of TTS system should be undertaken level wise. In the basic level methods like MOS for naturalness and MOS for intelligibility should be considered for evaluating synthesized synthetic speech. Once the satisfactory results are obtained from basic methods; testing should be continued using detailed analysis methods like comprehensibility test, segmental evaluation and prosody evaluation.

Scope of testing should be defined in due consideration of system supported features and its limitations. Test data should be designed with the aim to cover system specifications and to check the system behavior for its predefined limitations. Apart from speech quality evaluation of TTS product specific features must also be tested like voice selection, speed up & speed down, pitch.

CHAPTER – II : Test Data

1. Introduction

Test data creation is the crucial part in testing lifecycle. Effectiveness of testing increases if the depth and breadth of test data is wisely designed. Depth is the amount of data used in test execution and if data set is too small it will be difficult to simulate the real life conditions and if it is too huge then it will be cumbersome to execute and manage the test.

Breadth is the extent of variations in test data. Mere by increasing the depth of test data one cannot assure that all variations are addressed properly. If all the variations are not well taken then there is a possibility that many test scenarios may left unchecked. Along with data variations, samples of real time data should also be taken. For e.g.: In Railway announcement system data variations must include train numbers, station name & codes, date and time.

Negative test cases to check system's behavior and its ability to handle unexpected input should also be included. For e.g.: In monolingual TTS, foreign words input should be given to the application.

Prerequisite of data collection is the awareness about the application under test which includes features, supported languages, application area and its limitations.

2. Source of Data (web based & manual data typing)

Primary source of data harvesting is web and the data of not so popular languages is also available here. But if the data is not adequate enough to cover all test scenarios and data variations then other source should be opted. Another source is printed material in form of books, magazines, newspaper through which one can refer and get it typed.

There are few test scenarios for which one may not get data from above mentioned sources; synthetic data creation is the only mechanism to get such cases covered despite being time consuming and expensive. In TTS, data for testing language specific features one needs to type it manually.

Meta data of the collected data should be designed which will help in analyzing the collected data and for future reference.

3. Types of test data

This set comprises minimum variations of data which should be used for testing, irrespective to the domain of application under test.

1. **Akshara level data :**

   i) Varnamala : क, ख, थ, त

   kə, kʰə, tʰə, tə

   ii) Barakhadi : रि, पा, कै, बौ

   ɾi, paː, kæ, bɐɔː

2. Word level data

i. Difficult words :

E.g. : यूक्तयाभास, इक्यावन, दिक्स्थापन.

juktəjaːbʰaːs, ikjaːʊən, dikstʰaːpən

ii. MRT :

E.g. : कनक, जनक, सनक,ऐनक

3. Sentence level data

Following variations can be included in sentence level data:

i. Numerals – numerals can be read in any of the two manners.

a. Digit by digit number reading:

1. Phone no. : रमेश का नया मोबाइल नंबर 9970123456 है।
2. Address: ३०/११ हौज़ खास मार्किट नई दिल्ली ११०००३।
3. Aplha numeric: 11C, W3C.

b. Quantitative number reading:

1. Temperature: 30°C , -200°F
2. भारत में 1401 बाग़ ही बचे रह गए हैं ।
3. सन २००८ में सरकारी खजाने को 90 करोड़ रुपए का नुकसान हुआ था।
4. 10th (tenth) , X(tenth)

ii. Dates :

14 सितंबर 2013, 17-12-1989, 1934 ई., १० मार्च २०१३, 15/08/1947, १ .५ .२०१४.

iii. Abbreviations :

PWD, NDA, UPA, भाजपा, प्र. लि., govt., 100km,1000m,२min .

Sometimes expansion of abbreviation is context dependent. For e.g. CBI can be expanded as '*central bureau of investigation*' or 'central *bank of India'.*

iv. Salutations –

डॉ कुमार, श्री अशोक , प्रॉ ओम, कु. लता

v. Punctuations

{ . , ; : - _ ‘ “ ? ! () [] {} }

vi. Special symbols

List of symbols to be covered (@, #, $, %, *, |, \, /,<, >, +, =, -, ^, ~, `, degree symbol, rupee symbol)

vii. English words written in Latin and Indian scripts

1.सीडी install करना बहुत ही आसान कार्य है|

2.इस बार मेरे सब्जेक्ट्स हिस्ट्री और जियोग्राफी है।

viii. Semantically unpredictable sentences

1. आज बनाए बिल में चार हाथी घबराए।
2. तालाब दिखा नारंगी फूल गुलाबी गुड्डे।

4. Paragraph level data:

E.g. passage for MOS

हिन्दी के विकास में पहले साधु-संत एवं धार्मिक नेताओं का महत्वपूर्ण योगदान रहा। उसके बाद हिन्दी पत्रकारिता एवं स्वतंत्रता संग्राम से बहुत मदद मिली; फिर बंबइया फिल्मों से सहायता मिली और अब इलेक्ट्रॉनिक मीडिया के कारण हिन्दी समझने-बोलने वालों की संख्या में बहुत अधिक वृद्धि हुई है।

E.g. passage for Comprehensibility

अजी सुनते हो, मेरा पेट बहुत निकलता आ रहा है। शीशे में अपना शरीर देखती हूं, तो बड़ा खराब लगता है। सोचती हूं कि कोई बढ़िया जिम ज्वाइन कर लूं। पत्नी अपनी बाहों का हार डालते हुए पति से बोली। जिम मतलब... हजार रुपए महीना... मैं अफोर्ड नहीं कर सकता, वैसे ही इतने खर्चे हैं, अब तो अर्पित के

साथ आकांक्षा भी ट्यूशन पढ़ने जाने लगी है। पति अपनी असमर्थता जाहिर करते हुए समझाने लगा, फिलहाल मेरा बजट अभी इसकी इजाजत नहीं देता।

परेशान न हो, कुछ भी फर्क नहीं पड़ेगा मेरे पास इसका समाधान है। इतना कहकर पत्नी मुस्कराकर पति के कान में कुछ फुसफुसाने लगी। दूसरे दिन काम वाली की छुट्टी कर दी गई और तीसरे दिन गांव से बूढ़ी मां को फिर वापस बुला लिया गया।

Question to be asked from evaluator:

1. बच्चो के नाम बताइये?
2. बजट में कितने रुपये का फर्क पड़ता है?
3. माँ को गाँव से क्यों बुला लिया गया ?
4. कहानी का शीर्षक बताइए ?
5. कामवाली का क्या नाम था ?

4. Data set for prosody evaluation

Prosody Evaluation –

1. Types of sentences to check intonation

a. Interrogative: आपको कौन - कौन से फूल पसंद हैं?

b. Declarative : कल मैं दिल्ली जा रहा हूँ |

c. Exclamatory: अहा! कितना सुन्दर उपवन है।

d. Imperative : Order: खाना पकाओ। Request : कृपया शांति बनाये रखें।

e. Negative : यह मेरी पुस्तक नहीं है।

f. Emphatic : मैं दवाई लेता तो हूँ|

2. Sentences for analysing emotional prosody 5 basic emotions (anger, fear, jealous, joy, sad) will be checked.

g. Anger : तुमने काम समय रेहते क्यूं नहीं किया |

h. Fear: मुझे अंधेरे में जाने से डर लगता है |

i. Jealous: मेरी फोटो तुझसे ज़्यादा अच्छी है |

j. Joy: हम जीत गये |

k. Sad: लम्बी बीमारी के कारण सुभाष की मृत्यु हो गयी |

5. Domain specific data set

For testing domain specific application, test data should include vocabulary of targeted domain. It is important to check the performance of the application for domain specific data set, as it will help in inspecting the behavior of the application in real world scenario. While collecting test data awareness of domain is important for assuring its correctness and completeness.

This set will comprise all the keywords, sentences related to targeted domain. For instance in case of agriculture domain test data must include names of seeds, mandi, farming techniques, crops, pesticides, fertilizers, mechanism of cultivation, tools, type of soil, agriculture related queries.

6. Example applications with recommended test data suite

Application	Screen reader	Announcement System	PDF Reader	Domain specific TTS	Navigation app
Target Users	People with disability	All	All	Domain specific user	User of handheld devices and automobiles
Description	For ease of access to digital content	Automated announcement system	In spite of being busy in any work, one can read with their ear using pdf reader.	Domain specific TTS can be made available as web plug-in or mobile based app. It will provide information of specific domain via synthetic voice.	It directs by speaking the route thereby finding the destination, considerably useful while driving.
Test Data	1. Different types of data. 2.Menu, icons.	1. Different types of data. 2. Emphasis on numbers, named entities. 3.Application specific keywords like delay, departure, arrival.	1.Common data set 2.product specific features like comments, highlighted text.	1. Different types of data. 2. Emphasis on numbers, named entities. 3. For tourism domain, keywords like guesthouse, destination, विरासत, मार्ग.	1. Different types of data. 2 Emphasis on numbers, named entities, measurement. 3. Application specific keywords like direction,street.

7. Data Validation and Cleaning

Completeness of data is judged by its ability to satisfy the goal, so it is vital to meet the quality measures of data. If the errors in test data are neglected it can lead to misleading results thereby affecting the whole purpose of testing. Data Validation and Data cleaning is done to make the test data 'fit for use'.

In the data collected from any data collection techniques there is a possibility of errors. Errors can be classified in following types:

- Typo errors
- Spelling & grammar mistakes
- Data duplicity
- Data irrelevance with targeted domain
- Incomplete data(sentences/paragraphs)

To remove such errors, data collection should follow validation process. **Validation** identifies causes of error and prevents those errors from re-occurring. Validation should be performed by language experts with awareness about application domain.

Data cleaning is done by fixing the errors that are found in the validation.
There is a possibility that the corrected data also contain errors. So it is advisable to retain old and corrected data for further verification. Manual data cleaning is prone to errors and it is an expensive & time consuming job.

Handling large amount of data is a difficult task, a metadata should be prepared which will help in analyzing the data and this data can be referred in future also[15].

Annexure I: ABBREVIATIONS & ACRONYMS

This TTS Testing Strategy uses the following abbreviations:

TTS - Text to Speech

NVDA- Non Visual Desktop Access

SUS – Semantically Unpredictable Sentence

DRT- Diagnostic Rhyme Test

MRT- Modified Rhyme Test

CDAC – Centre for Development of Advanced Computing

GIST – Graphics and Intelligence based Script Technology

Annexure II: Template to identify scope of application under test

Support→		Others
Operating systems/platforms	Windows ☐ Linux ☐ Android ☐ Others ☐	
Delivery format	Desktop based ☐ Plug-in based (Browser based) ☐ Mobile integration ☐ Others ☐	
Domain of TTS (Like Tourism, Agriculture)		
Language supported	Monolingual ☐ Bi-lingual ☐ Multilingual ☐	
Frame work	Open source ☐ Free ☐ In-house ☐ Corporate (Licensed) ☐	

Audio data based format	MP3 ☐ .wav ☐ Others ☐	
Data base recorded environment	Noisy /field ☐ Studio/silent ☐	
TTS compatible	True Type ☐ Unicode ☐ UTF-8 ☐ Others ☐	
Speech generation process	Concatenation ☐ Parametric ☐ Articulatory ☐ Others ☐	
Speech output format	MP3 ☐ .wav ☐ Others ☐	
Sampling frequency and stereo/Mono bit rate		
Preprocessing		

Modules	Text Preprocessing ☐ Grapheme to phoneme ☐ Prosody ☐	
Speech database size Memory utilization		
Development environment Eg: windows 7 Visual studio 2008 C #		

REFERENCES

1. SRS of “Development of Text to speech system in Indian Languages ‘Phase-II’ ”.
2. ITU-T P 800 08/96 Methods for objective and subjective assessment of quality.
3. ITU-T P.85 1994 A Method For Subjective Performance Assessment Of The Quality Of Speech Voice Output Devices.
4. Paper from IIT Madras. Measuring Quality for Text-to-Speech Systems using degraded MOS scale and Word Error Rate by TTS Consortium.
5. Evaluation of MU-TALK Speech Synthesis System by Andrew Lampert.
6. A Common Attribute Based Unified HTS framework for speech synthesis in Indian languages - 8th ISCA Speech Synthesis Workshop
7. Testing Strategy of TTS phase I by CDAC Gist
8. The Blizzard Challenge 2013 – Indian Language Tasks
9. The Blizzard Challenge 2005: Evaluating corpus based speech synthesis on common datasets by Alan W Black and Keiichi Tokuda
10. Evaluation of TTS System in Intelligibility and Comprehension Tasks - Yu Yun Chang
11. Measuring speech quality for text-to-speech systems: Mahesh Viswanathan
12. ITU-T G.108.1 Guidance for assessing conversational speech quality
13. Quality Evaluation of Synthesized speech : By Vincent J. van Heuven
14. Evaluation Of A Multilingual TTS System With Respect To The Prosodic Quality by Rüdiger Hoffmann, Diane Hirschfeld
15. Principles of Data Quality by Aurther D Chapman.
16. Text to speech accuracy testing – 2003 by Voice Information Associates

Contributors to TTS Testing Strategy document:

1. Prof. Hema Murthy, IIT Madras (TTS Consortium Leader)
2. Prof. R.M.K Sinha, JSS Academy, Noida
3. Prof. A.G.Ramakrishanan., IISc Bangalore
4. Prof. S.S. Agrawal, KIIT Gurgaon
5. Shri. Mahesh D Kulkarni, CDAC GIST, Pune
6. Dr. Somnath Chandra, DeitY
7. Dr. Raimond Doctor, CDAC-GIST, Pune
8. Ms. Swaran Lata, DeitY
9. Ms. Tejaswini Patil, CDAC-GIST, Pune
10. Ms. Anusha Prakash, IIT Madras
11. Mr. Ankit Kesarwani, CDAC- GIST, Pune
12. Ms. Ruchi Raheja, CDAC GIST, Pune
13. Ms. Nisha Yadav, CDAC GIST, Pune