System Evaluation in Text-To-Speech by Statistical and Mean Opinion Score (MOS) Test

[1]Dr. Suhas Mache, [2]Dr. Sunil Nimbhore

[1]Department of Computer Science, R. B. Attal Arts, Science & Commerce College, Georai, Dist. Beed (MS)-India. suhas.mache@gmail.com

[2]Dept. of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS)-India. nimbhoress@gmail.com

***Abstract*— The aim of testing and evaluation of the TTS system is to determine the system performance and accuracy. Text-To-Speech (TTS) framework is to change over a subjective given content into a comparing talked waveform or fake creation of human discourse. This paper presents a TTS system for Pali language that uses concatenation using unit selection method. According to the standard measurement of speech and voice, our result analysis follows the standard measurement of speech and voice. The evaluation methods were designed to check the system accuracy and speech quality i.e. measuring the intelligibility of synthesized speech by Mean Opinion Score (MOS) test; Speaking Rate Test. The test result shows the overall accuracy of the TTS system is excellent and capable of generating natural sounding synthesized speech.**

Keywords*— *Text-To-Speech, unit selection, speech quality, intelligibility, MOS

I. Introduction

Text to speech system (TTS) converts text into voice using a speech synthesizer it is the artificial creation of human dialogue [1]. In recent years a lot of research is going on speech synthesis. Speech plays important role in day to day life communication. Speech synthesis i.e. Text-To-Speech is the method of converting the written content into machine-generated artificial speech [2]. We have selected concatenative unit selection method to develop Text-To-Speech (TTS) synthesis for Pali language. Concatenative speech synthesis systems read a text and render into speech by joining pre-recorded speech units to each other [3]. The Unit selection based corpus method is bulky corpus methods use to select the speech units and concatenate.

We have designed and developed an intelligible and natural sounding corpus-based concatenative speech synthesis system for Pali language. The implemented system is divided into two sections the front-end deals with text processing [1] and back end speech generation. The inputted text is first analyzed, normalized and transcribed into a phonetic representation [12]. The unit selection algorithm is based on the best path in the network of the units [4]. The second section back-end of the system is responsible for speech waveform generation. In this work, the different unit sizes such as vowels, consonants, syllables, digits, and words have experimented. In unit selection based concatenative speech synthesis, joint cost also known as concatenative cost, which measures how well two units can be joined together [5][6].

II. Methodology

The aim of testing and evaluation of the TTS system is to determine the system performance and accuracy. It also used to judge the speech quality in terms of its similarity to the human voice and by its ability to be understood.

A. *Test Data*

Test speech data plays a vital role in testing process and it effect on overall test outcomes. Test data is satisfactory sufficient to cover all the functionalities of the system under test. Different functionalities of the TTS system can be evaluated by investigative in general output speech. It should be designed in such a way that it covers all possible variations including numerals, vowels, consonants, words and connected words.

The Text-to-Speech system is evaluated by three different methods i.e. Objective Test, Subjective Test and Acoustic Measurements of speech. The objective test contains

DOI : 10.18231/2454-9150.2018.0879

Accuracy Test and subjective test intelligibility test by mean opinion score [7].

B. *Accuracy Test*

To conduct accuracy tests proper selection of test data is important. All such data whose predictable output is well defined can be measured for accuracy test [8]. For Accuracy measure, it just checks the pronunciation of total correct data such as numerals (digits), vowels, consonant and words with a total number of the text of the above input. The formula is

$$Accuracy = \frac{No.of\ Correct\ Pronounced\ Speech\ Data}{Total\ No.of\ Text\ Data}\ X\ {:}100$$

C. *Subjective Evaluation Metrics*

Intelligibility test by Mean Opinion Score (MOS)

The effective performance of a Text-to-Speech synthesis system can be properly measured by conducting subjective listening tests [8]. A mean opinion score (MOS) test was conducted. MOS is the arithmetic mean of all the individual scores and it gives the numerical indication of the perceived speech quality. To check the intelligibility of synthesized speech. As the part of this evaluation, we selected 10 (ten) sentences and 10 listeners. The listeners were asked to give a rating from 1 to 5 to each utterance. The definition of the rating is shown in table 1.

Table 1 Intelligibility by MOS Scale

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

III. RESULTS

To evaluate the system we have tested all possible variations including numerals, vowels, consonants, words and connected words.

i) System Accuracy Test

Numerals

$$Accuracy = \frac{100}{100}\ X\ 100 = 100\% \qquad (1)$$

Vowels

$$Accuracy = \frac{8}{8}\ X\ 100 = 100\% \qquad (2)$$

Consonants

$$Accuracy = \frac{32}{32}\ X\ 100 = 100\% \qquad (3)$$

Syllables

$$Accuracy = \frac{341}{341}\ X\ 100 = 100\% \qquad (4)$$

Words

$$Accuracy = \frac{71}{100}\ X\ 100 = 71\ \% \qquad (5)$$

Connected words

$$Accuracy = \frac{42}{100}\ X\ 100 = 42\ \% \qquad (6)$$

ANN Words

$$Accuracy = \frac{68}{100}\ X\ 100 = 68\% \qquad (7)$$

A. *Overall Performance of the system*

The overall TTS-System performance is computed by calculating the percentage of correct phonemes (i.e. consonants and vowels), 1 to 100 digits, short words, connected words, and ANN trained connected words of Pali language.

Table 2 Overall System Test Results

Test	Type of Data	Accuracy (%)
1	Vowels	100 %
2	Consonants	100 %
3	Syllables	100 %
4	Digits (1 – 100)	100 %
5	Short words	71 %
6	Connected words	42 %
7	ANN trained words	68 %
Average		**83. 00 %**

All these tests show that the accuracy of the developed TTS system is 83.00 %.

B. *Subjective evaluation (Listing Tests MOS)*

The effective performance of a Text-to-Speech synthesis system in terms of similarity with human voice can be properly measured by conducting subjective listening tests [9]. i.e. Mean Opinion Score (MOS) test was conducted. While objective measures are useful in comparing detailed system characteristics, the effective performance of a Text-to-Speech synthesis system can be properly measured by conducting subjective listening tests [9]. This test finding the relationships between intelligibility and comprehensibility in speech synthesizers and tries to design an appropriate comprehension task for evaluating the speech synthesizers' comprehensibility [10,11]. A mean opinion score (MOS) test was conducted. MOS is the arithmetic mean of all the individual scores and it gives the numerical indication of the perceived speech quality. To check the intangibility of synthesized speech. A part of this evaluation, we selected 10 (ten) sentences and 10 listeners to check the quality of speech and give rating from 1 to 5 to each utterance. The definition of rating was 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent. Table 3 shows the Scale of Mean Opinion Score and Table 3 shows Mean Opinion Score (MOS) Test result.

Table 3 Mean Opinion Score Test (MOS)

Sentence	L 1	L 2	L 3	L 4	L 5	L 6	L 7	L 8	L 9	L 10	MOS
S_1	4	3	4	4	5	4	5	5	4	4	4.2
S_2	3	4	4	4	3	4	4	4	4	4	3.8
S_3	4	4	4	3	4	5	4	4	3	5	4
S_4	4	4	4	4	4	4	3	4	4	5	4
S_5	4	3	3	5	4	4	5	3	4	5	4
S_6	5	4	5	4	5	5	4	5	5	5	4.7
S_7	4	3	4	4	4	4	3	3	4	4	3.7
S_8	3	4	3	4	3	4	3	4	3	4	3.5
S_9	4	3	4	3	4	4	4	3	4	4	3.7
S_10	4	4	3	4	3	3	3	4	4	4	3.6
Average											**3.92**

In the above table, L1 to L10 are ten different listeners and S_1 to S_10 are ten different sentences. The above MOS test shows average result 3.92, here we conclude that synthesized speech near to good intangibility.

c. Speaking Rate Test

The rate of speaking can be defined as the number of syllables or words speak by system per second. The average duration of a syllable is around 250 – 300 ms i.e. 2 – 3 syllables per second. If system's speaking rate is higher than this range it becomes sloppy speech and slower than normal range leads to elongation of duration of syllables. The table 4 shows speaking rate of system.

Table 4 Speaking Rate Test

Type of data	Speaking Rate
Syllable	2 -3 / second
Word (short)	1 word / 0.7 second
Word (long)	1 word / second

IV. CONCLUSION

The evaluation methods were designed to check the system accuracy and speech quality. The evaluation has been done at several levels, such as digits, vowels, consonants and words level. The overall test result shows the accuracy of the developed Text-To-Speech system is 83%. Here we have concluded that the Text to Speech conversion provides very good accuracy.

A subjective listing test for measuring the intelligibility of synthesized speech by Mean Opinion Score (MOS) test was also conducted. The MOS test gives 3.92 scores; this numerical indication shows the perceived speech quality is in a good range. The Speaking Rate Test also has been conducted. The test shows the rate of speaking in the number of syllables or words produced (spoken) by the system per second.

REFERENCES

[1] Mache, Suhas R., Manasi R. Baheti, and C. Namrata Mahender. "Review on text-to-speech synthesizer." International Journal of Advanced Research in Computer and Communication Engineering 4.8 (2015): 540.

[2] Dutoit, Thierry. An introduction to text-to-speech synthesis. Vol. 3. Springer Science & Business Media, 1997.

[3] Tatham, Mark, and Katherine Morton. Developments in speech synthesis. John Wiley & Sons, 2005.

[4] Black, Alan W., and Nick Campbell. "Optimising selection of units from speech databases for concatenative synthesis." (1995).

[5] Vepa, Jithendra, and Simon King. "Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis." IEEE Transactions on audio, speech, and language processing 14.5 (2006): 1763-1771.

[6] Vepa, Jithendra, and Simon King. "Subjective evaluation of join cost and smoothing methods." (2004).

[7] Klatt, Dennis H. "Review of text-to-speech conversion for English." The Journal of the Acoustical Society of America82.3 (1987): 737-793.

[8] TDIL, Text to Speech Testing Strategy Version 2.1 (2014) pp. 1-46

[9] Rosenberg, Andrew, and Bhuvana Ramabhadran. "Bias and statistical significance in evaluating speech synthesis with mean opinion scores." In Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017), Stockholm, Sweden, pp. 20-24. 2017.

[10] Chang, Yu-Yun. "Evaluation of TTS systems in intelligibility and comprehension tasks." proceedings of the 23rd Conference on Computational Linguistics and Speech Processing. Association for Computational Linguistics, 2011.

[11] Sunil S. Nimbhore, Rakesh J. Ramteke "Implementation of English-Text to Marathi-Speech (ETMS) Synthesizer", at International Organization of Scientific, Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. VI (Jan – Feb. 2015), PP 34-43

[12] Mache, Suhas and C. Mahender, Namrata "Development of Text-to-Speech Synthesizer for Pali Language", in IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 35-42.

DOI : 10.18231/2454-9150.2018.0879