Volume 2, Issue 1, January 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

EVALUATION OF KANNADA TEXT-TO-SPEECH [KTTS] SYSTEM

D.J.Ravi[1]
Department of Electronics and Communication
[1]Research Scholar, JSS Research Foundation
Mysore, India.

Sudarshan Patilkulkarni[2]
Department of Electronics and Communication
[2] Assistant Professor, SJ College of Engg.,
Mysore, India.

***Abstract*— The paper discusses evaluation tests namely Mean Opinion Score (MOS) test - carried out to examine the naturalness of the synthesized output of Kannada Text-To-Speech [KTTS] system, Diagnostic Rhyme Test (DRT) and Comprehension Test (CT) carried out to measure the intelligibility of KTTS system. A comparative study between recorded and synthesized speech carried out using two techniques namely DTW (Dynamic Time Warping) and Linear Predictor Co-efficient (LPC) Spectral Distance to measure how close the synthesized output is to a naturally spoken phrase. The tests helped in concluding that the synthesized speech output of KTTS is almost natural.**

Keywords-component; Mean Opinion Score (MOS); Diagnostic Rhyme Test (DRT); Comprehension Test (CT); Dynamic Time Warping (DWT); Linear Predictor Co-efficient (LPC); Kannada Text-To-Speech [KTTS] System.

I. EVALUATION TESTS

The basic criteria for measuring the performance of a TTS system can be listed as the similarity to the human voice (*naturalness*) and the ability to be understood (*intelligibility*). The ideal speech synthesizer is both natural and intelligible, or at least try to maximize both characteristics. Therefore, the aim of TTS is also determined as to synthesize the speeches in accordance with natural human speech and clarify the sounds as much as possible. For overall quality evaluation, the International Telecommunication Union (ITU) recommends a specific method that, in this author's opinion, are suitable for also testing *naturalness* (ITU-T Recommendation P.85 1994) [1,2]. Several methods have been developed to evaluate the overall quality or acceptability of synthetic speech [3,4]. Diagnostic Rhyme Test (DRT), Comprehension Test (CT) and Mean Opinion Score (MOS) are the most frequently used techniques for the evaluation of the naturalness and the intelligibility of TTS systems. Naturalness and intelligibility of the Kannada TTS system is tested by MOS and CT-DRT respectively.

II. MEAN OPINION SCORE [MOS]

The study was tested by making use of the Mean Opinion Score (MOS). The MOS that is expressed as a single number in the range 1 to 5, where 1 is lowest perceived quality and 5 is the highest perceived quality [3]. MOS tests for voice are specified by ITU-T recommendation. The MOS is generated by averaging the results of a set of standard, subjective tests where a number of listeners rate the perceived audio quality of test sentences read aloud by both male and female speakers over the communications medium being tested. A listener is required to give each sentence a rating using the rating scheme in Table I. The perceptual score of the method MOS is calculated by taking the mean of the all scores of each sentence.

TABLE I. MEAN OPINION SCORE

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

In the context of this study, 10 sentences that are provided in Table II are used for tests and 25 native Kannada speakers employed in the evaluation. For each sentence, MOS ranges that are assigned by the listeners are shown in Table III. Distribution of MOS values for test sentences by testers and average MOS values for each sentence are given in Figure 1 and Figure 2 respectively.

TABLE II. TEST SENTENCES

Sent	Kannada	English
S1	ನಿನ್ನ ಹೆಸರು ಏನು?	What is your name?
S2	ಹೂ ಎಷ್ಟು ಚೆನ್ನಾಗಿದೆ	What a beautiful flower
S3	ನನಗೆ ಕೆಲಸ ಸಿಕ್ಕಿತು	I got a job
S4	ಏಕೆ ಹೀಗೆ ಮಾಡಿದೆ?	Why did you do this?
S5	ನನಗೆ ತುಂಬಾ ಬೇಜರಾಗಿದೆ	I am very unhappy
S6	ಸೀತೆ ಯಾವ ಕಡೆ ಹೋದಳು?	Where did Sita go?
S7	ಭಾರತ ನನ್ನ ದೇಶ	Bharath is my Country
S8	ಇಂದು ಯಾವ ದಿನ?	What day is today?
S9	ಕೆಂಪು ಹೂ	Red flower
S10	ನಮ್ಮೂರು ಮೈಸೂರು	I am from Mysoru

TABLE III. MOS RANGE AND SCORES FOR EACH SENTENCE

Sent. N	Excellent 5	Good 4	Fair 3	Poor 2	Bad 1	Sum Listener	Point Sum	Avg.
S1	10	13	2	-	-	25	108	4.32
S2	09	15	1	-	-	25	108	4.32
S3	09	11	3	1	1	25	101	4.04
S4	10	11	3	1	-	25	105	4.20
S5	10	13	1	1	-	25	107	4.28
S6	08	14	2	1	-	25	104	4.16
S7	09	13	2	1	-	25	105	4.20
S8	10	12	3	-	-	25	107	4.28
S9	07	13	3	1	1	25	99	3.96
S10	08	12	3	2	-	25	101	4.04
Sum	**90**	**127**	**23**	**8**	**2**	**250**	**1045**	**4.18**

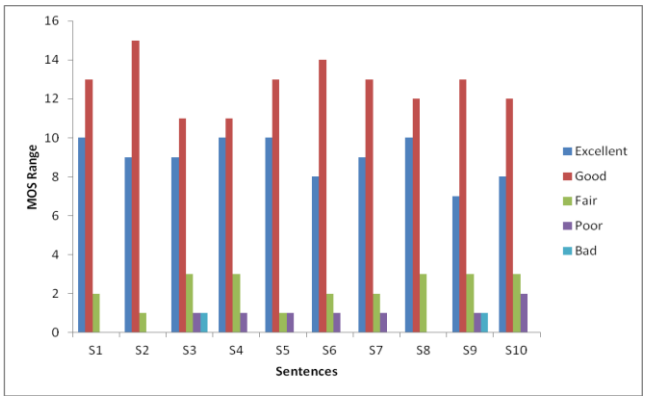

Figure 1. Distribution of MOS values for test sentences assigned by testers

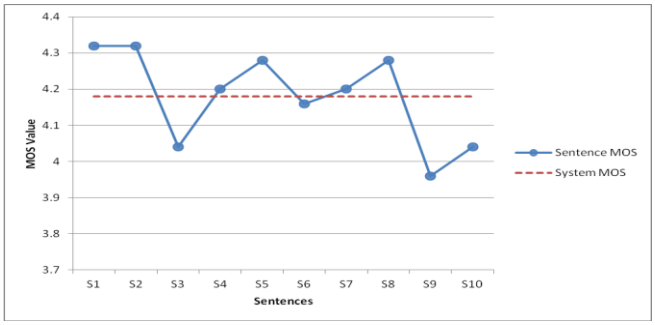

Figure 2. Distribution Average MOS values for each sentence and system average

III. MEAN OPINION SCORE OF SYNTHESIZED EMOTIONAL SPEECH

Quality of the speech is subjective in nature, speech which appears good to one person may not appear good to other, and hence it was decided to collect Mean Opinion Score (MOS) about the speech quality [3]. Some sentences neutral and emotional were synthesized. These synthesized speech were made to hear by 25 randomly selected people. The method of synthesis was not revealed to them, they were asked to assign a score on the scale of 0 to 10 for neutral as well as emotional speech. The results are tabulated in Table IV. Distribution of MOS percentage for Test Sentences in different emotions is shown in Figure 3. These results show that the synthesized speech quality is very high.

TABLE IV. EMOTION RECOGNITION RATE OF SYNTHESIZED SPEECH IN % (MEAN OPINION SCORE)

Score	Neutral	Anger	Sadness	Happiness
Poor(0 – 3)	5%	15%	10%	15%
Good(4 – 7)	25%	60%	60%	65%
Very Good(8-10)	70%	25%	30%	20%

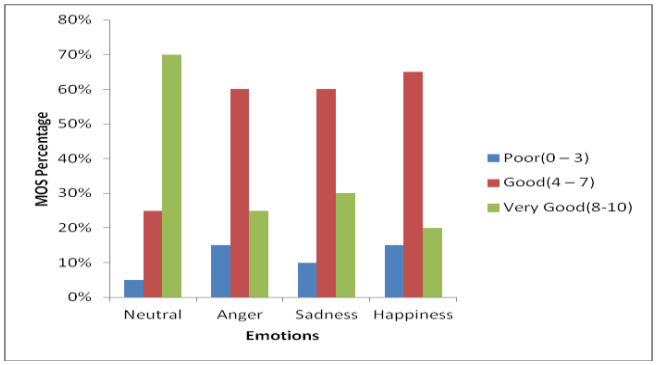

Figure 3. Distribution of MOS percentage for Test Sentences in different Emotions

IV. COMPREHENSION TEST

In the comprehension tests, a subject hears a few sentences or paragraphs and answers the questions about the content of the text, so some of the items may be missed [4]. It is not important to recognize one single phoneme, but the intended meaning. If the meaning of the sentence is understood, the 100% segmental intelligibility is not crucial for text comprehension and sometimes even long sections may be missed.

In the comprehension tests three subtests are applied. In all three cases, the testers are allowed to listen to the sentences twice. In the majority of the tests, success is achieved for the first listening trial, and second one also improves the results.

First comprehension subtest has 5 sentences and 5 questions that are shown Table V about the content. Listeners answer the question about content of each sentence. First listening trial accuracy is calculated as the ratio of number of correct answers given by the testers to the whole set of correct answer as (T/N=118/125) =0.944 and second trial accuracy is obtained as 1.

Second subtest is about answering common questions. It contains 5 sentences, S1-S5. Listeners answer the questions. Question sentences and number of correct answers at first and second listening are shown in Table VI. The results indicate that the understandability of the system is very high. Additionally, the accuracies of the first and second listening trials are 1.

Third subtest is applied as a filling in the blanks test. There are 5 noun phrases, one word of the phrase is provided to the listener and other is left as blank. The testers listened to the speech and filled in the blanks. Noun phrases and number of correct answers at first and second listening trial are shown in Table VII. The words that are underlined and given in capital letters are the blanks in the test. The accuracies of the first and second trial are 0.96 and 1, respectively by achieving a high understandability rate.

TABLE V. LISTENING COMPREHENSION

Sent. N.	Content question (Kannada)	Content question (English)	1st	Acc	2nd	Acc
S1	ನಿನ್ನ ವಯಸ್ಸು ಎಷ್ಟು?	How old are you?	24	0.96	1	1.00
S2	ಅವರು ಏನು ಮಾಡಿದರು?	What did they do?	23	0.92	2	1.00
S3	ಯಾರು ಬಂದರು?	Who came?	25	1.00	-	1.00
S4	ರಾಮ ಎಲ್ಲಿಗೆ ಹೋದ?	Where did Rama go?	24	0.96	1	1.00
S5	ನೀವು ಹೇಗಿದ್ದೀರಾ?	How are you doing?	22	0.88	3	1.00

TABLE VI. ANSWERING COMMON QUESTIONS

Sent. N.	Kannada	1st	Acc	2nd	Acc
S1	ನಿನ್ನ ಹೆಸರು ಏನು?	25	1.00	-	1.00
S2	೨+೫=	25	1.00	-	1.00
S3	ಇಂದು ಯಾವ ದಿನ?	24	0.96	1	1.00
S4	ಕರ್ನಾಟಕದ ರಾಜಧಾನಿ ಯಾವುದು?	25	1.00	-	1.00
S5	ಇಂದಿನ ಸುದ್ದಿ ಏನು?	25	1.00	-	1.00

TABLE VII. FILLING IN THE BLANKS

Kannada	English	1st	Acc	2nd	Acc
ಕಪ್ಪು ಬಿಳುಪು	black **WHITE**	24	0.96	1	1.00
ಚುನಾವಣೆ ಫಲಿತಾಂಶ	**ELECTION** results	23	0.92	2	1.00
ಬಿಸಿ ತುಪ್ಪ	hot **GHEE**	25	1.00	-	1.00
ನೀಲಿ ಆಕಾಶ	**BLUE** sky	24	0.96	1	1.00
ಚಿನ್ನ ಬೆಳ್ಳಿ	gold **SILVER**	24	0.96	1	1.00

V. DIAGNOSTIC RHYME TEST

Diagnostic rhyme test (DRT) [4] is an American National Standards Institute (ANSI) standard for measuring speech intelligibility (ANSI S3.2-1989). DRT, introduced by Fairbanks in 1958, uses a set of isolated words to test for consonant intelligibility in initial position. DRT is used how the initial consonant is recognized properly. In the DRT test of the current system, the consonants that are similar to each other are selected and the listeners are asked to distinguish the correct consonant among the similar sounding alternatives [7]. The letters that have the same way out such as ‘b’ and ‘p’ are plosive and bilabial consonant and can be easily misunderstood.

The similar sounding words that are used for DRT and number of correct answers for the first and the second listening trials are shown in Table VIII. Bold words indicate the correct answers. Listeners choose one word from the table they hear. The accuracies are calculated above 0.93 and mostly close to 1 especially after the second trial.

The summary of accuracies of the CT and DRT tests are given in Table IX. It shows system intelligibility rate very high and satisfactory.

TABLE VIII. WORDS FOR DIAGNOSTIC RHYME TEST (DRT)

Recognize	Listening	word 1	word 2	Acc.
P - b		**ಪರಿ (Manner)**	ಬರಿ (Empty)	
	1st	22	3	0.88
	2nd	24	1	0.96
t - d		ತಳ (Bottom)	**ದಳ (Petal)**	
	1st	2	23	0.92
	2nd	-	25	1.00
ṭ - ḍ		**ಟಗರು (Goat)**	ಡಮರು (Drum)	
	1st	24	1	0.96
	2nd	25	-	1.00
k - g		**ಕವಿ (Poet)**	ಗವಿ (Cave)	
	1st	25	-	1.00
	2nd	25	-	1.00
m - n		**ಮನೆ (House)**	ನೆನೆ (Remember)	
	1st	23	2	0.92
	2nd	25	-	1.00

TABLE IX. COMPREHENSION TESTS AND DIAGNOSTIC RHYME TEST (DRT) ACCURACY

Tests	Listening Trial Accuracies	
	1st	2nd
Comprehension Test (CT)		
1. Answer the questions about the content	0.944	1.00
2. Answering common question	1.00	1.00
3. Fillings the blanks	0.96	1.00
Diagnostic Rhyme Test (DRT)	0.936	1.00

VI. Recorded versus Synthesized speech – A Comparative Study

Analysis is one of the important tasks of any work. It does the evaluation of the work. Analysis helps the developer to know how effective the system is and also helps in understanding the flaws. The results of the analysis can lead the development of the work in a completely new way. Since speech quality is subjective in nature, absolute measurements cannot be made. But it is possible to measure some relative quantities, which indirectly shows the quality of synthesized speech [5,6].
Following are the two analysis methods, which we used in our work.

1. Dynamic Time Warping distance measure
2. Linear Predictor Co-efficient Spectral distance measure

Since, Speech quality measurement is relative; it becomes necessary to have a reference object. The reference object in our work is a directly recorded wave. The creator of the database records a phrase, which acts as our reference. The aim of the work is to produce a synthetic wave, which resembles the recorded wave perfectly. The quality of the synthesized wave is said to be very high, if it resembles the recorded wave more. Above mentioned two methods are used to determine resemblance of two waves.

A. Dynamic Time Warping (DTW) distance measure

A person cannot speak a word twice, exactly same. The rate of speech varies. But the synthesized speech rate does not vary no matter how many times it is synthesized. Finding out Euclidean distance between two waves, of which one is spoken very fast is not possible and unscientific. The Euclidean distance between two similar speech waves, of which one is relatively fast may show complete mismatch. Hence, before calculating distance, two waves must be aligned.

This technique is used to find out distances for phrase ಕದ ತೆರೆ. Calculating DTW distance for complete phrase requires a large amount of memory in general purpose computer. Hence, it is applied word by word. Table X shows the DTW distances obtained.

TABLE X. Results of DTW analysis

Word	Distance between Recorded wave and Synthesized
ಕದ	59
ತೆರೆ	65

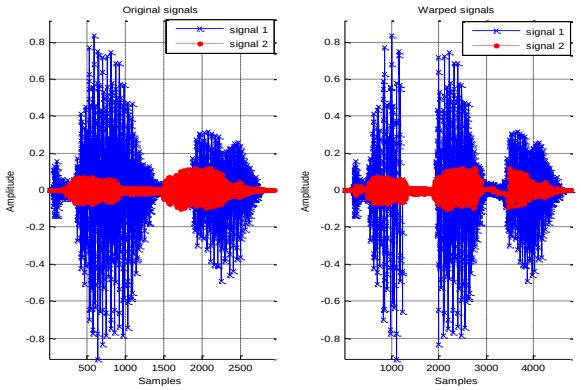

Figure 4. DTW algorithm for Reference and Synthesized wave

DTW algorithm is applied for measuring distance between Reference and Synthesized waves. Blue colored waves (Signal 1) represent synthesized waveforms. Before applying DTW algorithm the similarities are hidden. This is illustrated in the graph on the left side of Figure 4. After applying DTW all hidden similarities are visible and it is shown in right side of Figure 4. The distances are tabulated in Table X which clearly shows the distance between Reference and Synthesized are much lesser, which is the desired result. Similar analysis is made for other word ತೆರೆ.

B. Linear Predictor Co-efficient (LPC) Spectral Distance

Linear predictive coding (LPC) is a tool used in speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. It is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters. The DTW algorithm finds distance between two signals using time domain method. Only point to point distance is calculated in DTW. But in LPC spectral distance measurements, even frequency components are also taken care of. Similar to DTW analysis, LPC spectral envelope is calculated for Recorded Reference and Synthesize waves for complete phrase ಕದ ತೆರೆ. Then Euclidean distance is calculated these values are tabulated in Table XI. Figure 5 shows the Linear Predictor Spectral Estimation. The graphs for Recorded and Synthesized waves are almost same, this indicates the frequency contents are matching. The value in the Table 11 for synthesized wave is much nearer to 0.

TABLE XI. LPC spectral distance

Phrase	Distance between Recorded wave and Synthesized
ಕದ ತೆರೆ	0.1885

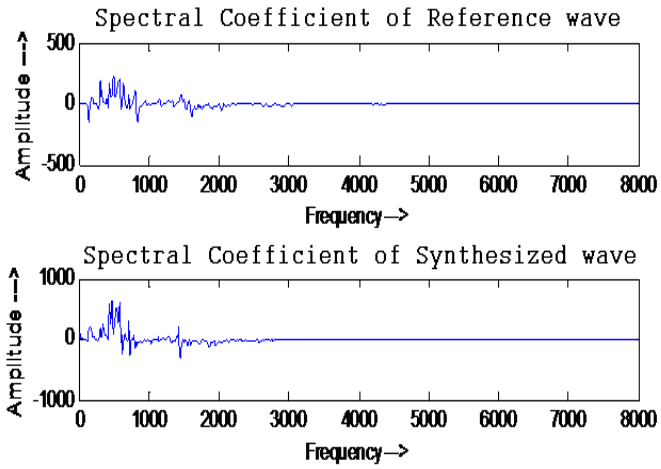

Figure 5. LPC Spectral estimation

VII. CONCLUSION

A number of subjective tests are used to measure the success of KTTS system. Naturalness defined as closeness to human speech and intelligibility defined as the ability to be understood are two measures used to evaluate the performance of any Text-to-speech system. Mean Opinion Score (MOS) test is carried out to examine the naturalness of the synthesized output of KTTS system. Diagnostic Rhyme Test (DRT) and Comprehension Test (CT) are carried out to measure the intelligibility of KTTS system. These evaluations have provided promising results. Further two techniques DTW (Dynamic Time Warping) and Linear Predictor Co-efficient (LPC) Spectral Distance are applied for comparison of synthesized output versus naturally recorded speech of a phrase. Comparison concludes that the distance between recorded (reference) and synthesized are much lesser, which is the desired result. These tests helped in concluding that the synthesized speech output of KTTS is almost natural.

REFERENCES

[1] Dmitry Sityaev, Katherine Knill and Tina Burrows, "Comparison of the ITU-T P.85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems", INTERSPEECH 2006 – ICSLP, pp. 1077-1080.

[2] Yolanda Vazquez Alvarez & Mark Huckvale, "The Reliability of the ITU-T P.85 Standard for the Evaluation of Text-To-Speech Systems". Department of Phonetics and Linguistics University College London, U.K., 2002, pp. 1-4.

[3] Mean opinion score[MOS] : wikipedia : http://en.wikipedia.org/wiki/Mean Opinion_Score

[4] Speech Quality and Evaluation: wikipedia : http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap10.html

[5] S´ergio Paulo, Lu´ıs C. Oliveira, "DTW-based Phonetic Alignment Using Multiple Acoustic Features", EUROSPEECH 2003 – GENEVA, pp. 309-312.

[6] Steven Bo Davis and Paul Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", 1980, pp. 195-216.

[7] B.B.Rajapurohit, "Acoustic Characteristics of Kannada", CIIL, Mysore.