RESEARCH ARTICLE OPEN ACCESS

Evaluation of Hidden Markov Model based Marathi Text-To-Speech Synthesis System

Monica Mundada*, Dr.Bharti Gawali**
**(Department of Computer Science & Information Technology Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India*
*** (Department of Computer Science & Information Technology Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India*

ABSTRACT
The objective of this paper is to evaluate the quality of HMM based Marathi TTS system. The main advantage of HMM technique is its ability to allow the variation in voice easily. The output speeches produced in this method have greater impact on emotion, style and intonation. The naturalness and intelligibility are the two important parameters to decide the quality of synthetic speech. Depending on the parameters specified the results of synthetic speech are categorized into 4 categories: natural speech, high quality synthetic speech, low quality synthetic speech and moderate quality synthetic speech. The results are obtained by using CT, DRT and MOS test.

Keywords: Mean Opinion Score (MOS), Hidden Markov Model (HMM), Diagnostic Rhyme test (DRT), Comprehension Test (CT), Naturalness and Intelligibility.

I. INTRODUCTION

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. The quality and evaluation of speech synthesis methods depends on the use of user end application to be built. For example reading machines for the blind, the speech intelligibility with high speech rate is usually more important feature than the naturalness. On the other hand, prosodic features and naturalness are essential when we are dealing with multimedia applications or electronic mail readers. The evaluation can also be made at several levels, such as phoneme, word or sentence level, depending what kind of information is needed. The evaluation procedure is usually done by subjective listening tests with response set of syllables, words, sentences, or with other questions. The test material is usually focused on consonants, because they are more problematic to synthesize than vowels. In Marathi language there are 12 vowels and 35 consonants [1].

Introduction of the paper should explain the nature of the problem, previous work, purpose, and the contribution of the paper. The contents of each section may be provided to understand easily about the paper.

II. HMM BASED SPEECH SYNTHESIS

Parametric representation of speech using HMMs is known as HMM based speech synthesis system. The advantage of using speech synthesis method based on HMMs is its ability to synthesize intelligible and natural sounding speech without requiring a huge training database. When speech is synthesized from HMMs directly, it is possible to synthesize speech with various voice characteristics [2]. Low memory requirements, flexibility and ease of adaptability to speaker‟s voice characteristics and speaking styles, are some of the factors that favored to choose the HMM based speech synthesis method over other methods. An HMM is a finite state machine which generates a sequence of discrete time observations. At each time unit i.e., frame, the HMM changes states according to state transition probability distribution and then generates an observation O t at time t according to the output probability distribution of the current state. Hence the HMM is a doubly stochastic random process model.

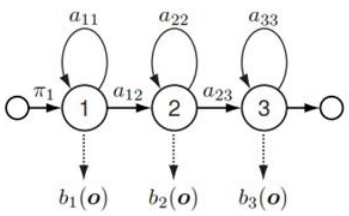

Fig 1. An example of HMM model

Figure 1 shows examples of HMM model in which the state index increases or stays the same as the time increases. One common HMM topology is to use three states. Each state is linked to the next state and back to itself again. This last transition is known as the self-transition probability and is basically the probability that the next observation is generated from the present state. A phone‟s state transition probabilities govern the durational characteristics of the phone; if the self-transition probabilities are high, it is more likely that more observations will be generated by that phone which means the overall phone length will be longer HMM based speech synthesis system mainly consists of two parts. One is the training part and the other is synthesis part. In the training part, spectrum and excitation parameters are extracted from speech database and modeled by phoneme HMMs. in this system, feature vector consists of spectrum and F0 parts. The spectrum part consists of Mel-cepstral coefficients, their dynamic features i.e. delta and delta-delta. F0 part consists of log F0, its delta and delta-delta. In the synthesis part, phonemes HMMs are concatenated according to the text to be synthesized. Then spectrum and excitation parameters are generated from the HMM [3]. The excitation generation module and synthesis filter module synthesize speech waveform using the generated excitation and spectrum parameters. The attraction of this approach is that voice characteristics of synthesized speech can easily be changed by transforming HMM parameters.

III. CHARACTERISTIC OF HMM BASED SPEECH SYNTHESIS

The HMM-based speech synthesis approach involves the training of HMM context-dependent HMMs & duration models. There are three factors which degrade the quality of synthesis system: vocoder, modeling accuracy, and over-smoothing. The synthesized speech by the HMM-based generation synthesis approach sounds buzzy since it is based on the vocoding technique. To alleviate this problem, a high quality vocoder such as multi-band excitation scheme have been integrated. The basic system uses ML-estimated HMMs as its acoustic models. Because this system generates speech parameters from its acoustic models, model accuracy highly affects the quality of synthesized speech. To improve its modeling accuracy, a number of advanced acoustic models and training frameworks have been investigated. In the basic system, the speech parameter generation algorithm is used to generate spectral and excitation parameters from HMMs. By taking account of constraints between the static and dynamic features, it can generate smooth speech parameter trajectories.

Advantages of the HMM-based generation synthesis approach are

- Its voice characteristics can be easily modified,
- It can be applied to various languages with little modification,
- A variety of speaking styles or emotional speech can be synthesized using the small amount of speech data,
- Techniques developed in ASR can be easily applied,
- Its footprint is relatively small.

IV. BUILDING THE CORPUS

The corpus created for the said research consist of highly professional male speaker. The total size of database consist of 100 SMS which are commonly used in day to day life. The SMS are designed which covers all the festivals, Love, Birthday wishes, Breakup messages, Great thoughts for inspiration, Funny SMS and special SMS like wedding anniversary, Praising someone etc. A noise less environment like a specially designed recording studio is required to avoid background noise while recording the speech files [4]. Recording is done at recording studio with high professional speaker. The distance from the microphone to mouth, speaking volume and speaking style is kept constant till the completion of recording. In this analysis the system is trained with the built up database. For testing the random sentence are designed which are out of database to check the intelligibility of actual working of TTS.

V. EXPERIMENTAL OBSERVATIONS

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. 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. The output of the synthesized wave files is processed for MOS test, CT test and DRT test. All the test are performed individually to check the intelligibility and naturalness associated with the speech signal. In this experiment 20 native Marathi speakers have been participated with 10 females and 10 males. The average age of participant is 21-23 years with undergraduate qualification.

5.1 Mean Opinion score (MOS) test

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, here 1 is lowest perceived quality and 5 is the highest perceived quality [5].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 1. In the context of this study, 10 random sentences are chosen for performing the test and 20 native Marathi speakers employed in the evaluation. For each sentence, MOS ranges that are assigned by the listeners are shown in Table 1.

Table 1. Mean Opinion Score

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

Table 2. MOS score given by listeners for each sentence

Sentence	Excellent	Good	Fair	Poor	Bad
S1	16	02	02	-	-
S2	17	02	1	-	-
S3	16	02	1	1	-
S4	17	02	1	-	-
S5	18	01	1	-	-
S6	17	02	1	-	-
S7	17	02	1	-	-
S8	16	02	1	1	-
S9	17	01	1	1	-
S10	18	01	1	-	-

From table 2 it is dissipated that out of 20 listeners , the 16 individuals scored the sentence 1 as excellent quality of speech, 02 listeners rated is as good and only 2 listeners rated it as fair. Like wise the calculation is performed for all 10 sentences. So it is drawn that 84.5% provides the excellent quality of speech, 8.5% provides the good rating speech where as fair quality speech is 5.5% and 1.5 % gives the poor quality of synthesized speech. The following table 3 shows the output of quality speech as follows. Distribution of MOS values for test sentences by testers and average MOS values for each sentence are given in Figure 2.

Table 3. Percentage evaluation of output speech.

Quality of synthesized speech	% percentage Evaluation
Natural quality Speech	84.5%
High quality synthetic speech	8.5%
Low quality synthetic speech	5.5%
Moderate quality synthetic speech	1 .5%

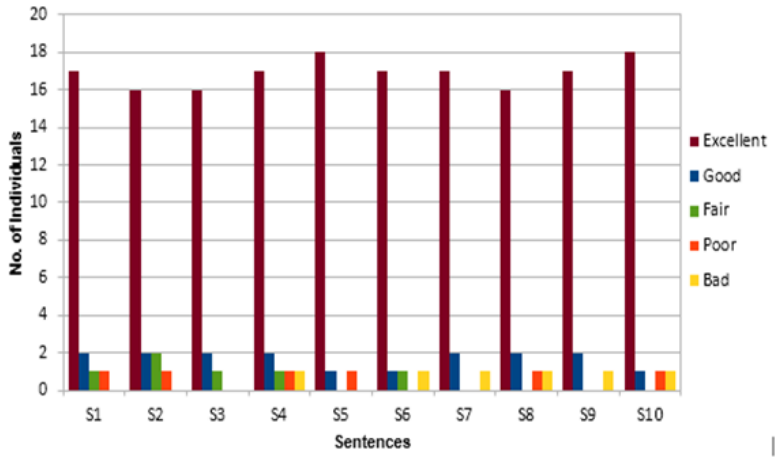

Figure 2. Distribution of MOS values for sentences tested by individual subject

5.2 Comprehension Test (CT)

This test is performed to check the degree of received messages being understood. In this test, initially the corpus of 10 minutes is made hear to participants. After that a set of 5 questions were asked based on the heard concepts. A 3-point scale (0, 1, 2) was applied in the experiment to score answers in the open-ended questions. If the responses to the comprehension questions were judged to be incorrect, 0 points were earned; if part of the answers were correct or the answers were too general and nonspecific, yet not wrong, 1 point would be given; and 2 points were given to the responses with fully correct and specific answers. A total of 10 points for 5 open-ended questions was assigned.

Table 4. Example for performing Comprehension test (CT)

Open-ended Question	Correct Answer	Listener Response	Score
सुविचार किती होते ?	05	05	02
		05	02
		04	01
		04	01
		Don't know	0

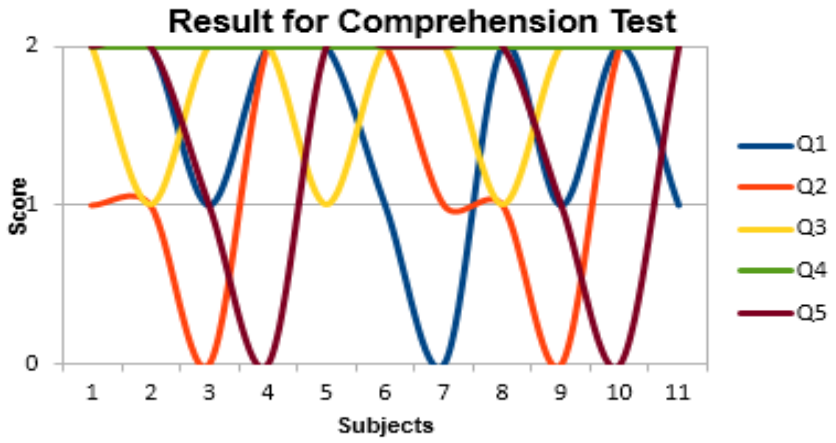

Figure 3. Line graph showing score for each question

The figure 3 shows the line graph for each individual subject with the respective score for five set of questions which are labeled as Q1, Q2,Q3,Q4 and Q5.The average score for each question set is 2 which is maximum allotted by the subject. Thus result indicates that the synthetic speech produced by the system is acceptable and understandable.

5.3 Diagnostic Rhyme Test (DRT)

This test is used to evaluate the intelligibility of TTS synthesis. The test uses monosyllabic words having a consonant-vowel-consonant pattern and it measures the capability of discrimination of the initial consonants for the system under consideration [6]. Listeners listen to monosyllabic words which differ only in the first consonant and have to choose which word they have heard from pairs (for example, naav/gaav, kiti/bhiti) This test is performed for testing one pair of monosyllabic word at a time. This evaluates the correct pronunciation and synthesis of words. The test is performed in between 10 users for 20 words. The 95% of words are correctly identified by the individual in this test.

VI. CONCLUSION

A number of subjective tests are used to measure the success of HMM based Marathi TTS 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. With regards to the definition of TTS, the random sentences are selected for performing the various test and evaluation. Mean Opinion Score (MOS) test is carried out to examine the naturalness of the synthesized output of Marathi TTS system. This test proves that the listeners have built the confidence in the naturalness of the voice produced. Comprehension Test (CT) and Diagnostic Rhyme Test (DRT) is performed to identify the intelligibility of the system. In DRT, almost most of monosyllabic words are identified correctly by the individual subject. These tests helps in concluding that the synthesized speech output of HMM based Marathi TTS system is almost natural.

The figure 3 shows the line graph for each individual subject with the respective score for five set of questions which are labeled as Q1, Q2,Q3,Q4 and Q5.The average score for each question set is 2 which is maximum allotted by the subject. Thus result indicates that the synthetic speech produced by the system is acceptable and understandable.

REFERENCES

[1]. Mrinalini Mukund Ghatage . "Pronunciation Problems of the Marathi Speakers "ISSN 1930-2940. April 2013

[2]. Mohammed Waseem, C.N Sujatha, "Speech Synthesis System for Indian Accent using Festvox", Internationalj ournal of Scientific Engineering and Technology Research, ISSN 2319-8885 Vol.03,Issue.34 November-2014, Pages:6903-6911.

[3]. Monica Mundada, Dr.Bharti W.Gawali."Comparative analysis of HMM based Marathi TTS using English and Marathi Text font".IJMER .ISSN: 2249–6645 . July 2016
[4]. Newton, P.S.R. : Review of methods of Speech Synthesis, EE Dept., IIT Bombay(November 2011).
[5]. Raitio, Tuomo, et al. "HMM-based speech synthesis utilizing glottal inverse filtering." Audio, Speech, and Language Processing, IEEE Transactions on vol.19, no.1, 2011, pp. 153-165
[6]. Steven Bo Davis and Paul Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", 1980, pp. 195-216.