ISSN: 2230-9926

*Available online at http://www.journalijdr.com*

IJDR

*International Journal of Development Research*
*Vol. 07, Issue, 09, pp.15236-15239, September, 2017*

ORIGINAL RESEARCH ARTICLE

Open Access

# A SURVEY ON METHODS OF TTS AND VARIOUS TEST FOR EVALUATING THE QUALITY OF SYNTHESIZED SPEECH

**[*,1]Pravin M. Ghate and [2]Shirbahadurkar, S. D.**

[1]Ph.D scholar NIELIT Dr B.A.M.U Aurangabad, India
[2]Zeal College of Engineering & Research, Narhe, India

**ARTICLE INFO**

***Article History:***
Received 09th June, 2017
Received in revised form
24th July, 2017
Accepted 06th August, 2017
Published online 29th September, 2017

***Keywords:***

Speech synthesis, Syllable,
Text to speech (TTS),
Mean Opinion Score (MOS),
Modified Rhyme Test (MRT).

****Corresponding author:***

**ABSTRACT**

The main objective of the paper is give an ideal about different methods of the text to speech synthesis for quality of synthesized speech for Marathi Language. A system converting textual information into speech is usually called as Text to speech (TTS). Speech is surely the most natural mode of communication in human's community. [Quartier, 2002]. This information can be evaluated at several non-exclusive stages of description. At the acoustic level, speech is considered as a mechanical wave that is an oscillation of pressure. The main claims of Speech Processing can be categorized as follows, Speech Recognition, Speech Synthesis, Speaker Recognition, Voice Analysis, Speech Enhancement, and Speech coding [Dutoit, 1997]. Out of the many application in this paper, speech synthesis is one of the discussion topic, Speech Synthesis: In Speech Synthesis, also called Text-to-Speech (TTS), the goal is to produce the speech, typically expressed in terms of naturalness and intelligibility of the produced voice.

**Citation: Pravin M. Ghate and Shirbahadurkar, S. D., 2017.** "A survey on methods of TTS and various test for evaluating the quality of synthesized speech.", *International Journal of Development Research*, 7, (09), 15236-15239.

## INTRODUCTION

Marathi language usually use to spoken by the native people of Maharashtra. Marathi is the group of Indo-Aryan languages which are a part of the larger group of Indo-European languages, all of which can be traced back to a common root. Among the Indo-Aryan languages, Marathi is the southern-most language. Matathi languages originated from Sanskrit. From Sanskrit, three different Prakrit languages, was developed which is simpler in structure. These are 1. Saurseni 2. Magadhi and 3. Maharashtri. Word was formed using Vowels are combined with consonants in forming syllables which ultimately form a word. A syllable is a element of organization for a order of speech sounds. For example, the word Marathi is composed of three syllables: *Ma,*ra and thi. A syllable is typically made up of a syllable nucleus (most often a vowel) with optional initial and final margins (typically, consonants). Syllables are habitually considered the phonological "building blocks" of words. It gives the influence the rhythm of a language, and its prosody, and stress arrays of word. In this paper, a sensible effort is made collective framework for building synthesis systems for Marathi languages. Speech synthesis is the artificial production of human speech by using text. A computer system used for this purpose is called a Speech Synthesizer, and can be implemented using different software or hardware.

### Implementation method of text to speech Marathi and other language

**Articulatory Synthesis***:* In this method of speech synthesis, computer and its hardware are used to mimic humans like speech using text. In this system the human speech production model based on human vocal tract and the articulation processes occurring there. These model uses the vocal and nasal tracts are treated as tubes that are attached with closures for articulators such as the tongue, jaw, and lips. A arifical arrangement was done which is same like vocal tract and speech is created by digitally simulating the flow of air through the representation of the vocal tract. Due to mathematics concncept method is highly complex and still is commercially unsuccessful.

### Formant Synthesis

Formant synthesis employs some set of rules to synthesize speech using the formants that are the resonance frequencies of the vocal tract. Since the formants constitute the main frequencies that make sounds distinct, speech is synthesized using these estimated frequencies

### Concatenative Synthesis

In this model of speech synthesis concatenative speech synthesis generate most natural and intelligible synthesized speech, In this process of generating the sound from text , as it avoids the difficult problem of modeling human speech production. it requires a lot of memory for to store the speech and There are three main approaches for concatenative speech synthesis: Diphone synthesis, Domain specific synthesis and Unit selection synthesis.

### Different method of Implementation

For implementation of text to speech some of the problems may be solved with methods described below and the use of concatenative method is increasing due to better computer capabilities (Donovan 1996) [10].

### **Pitch Synchronous Overlap Add** (PSOLA) Method

The concept of Pitch Synchronous Overlap Add) method was originally developed at France telecom (CNET). Pitch is one of main parameter in the speech analysis. In this speech synthesis method prerecorded speech samples is used to generate synthesized speech, such as Pro Verbe and HADIFIX (Donovan 1996) [10].There are different types of the PSOLA algorithm and all of them work in essence the same way. Time-domain version, TD-PSOLA, is the most commonly used due to its computational efficiency (Kortekaas et al. 1997)[11]. A algorithm consist of three basic steps. 1. The analysis step 2. Separate frame 3.Ovelapping and adding phase. Below the mathematical formula for . Short term signals $x_m(n)$ are obtained from digital speech waveform $x(n)$ by multiplying the signal by a sequence of pitch-synchronous analysis window $h_m(n)$:

$$x_m\,(n) = h_m\,(t_m\, - n)x(n)$$

where $m$ is an index for the short-time signal.

### Linear Prediction based Methods

This is one of method originally proposed for speech coding systems, but same concept which may be also used in speech synthesis. The first speech synthesizers were developed from speech coders like formant synthesis, LPC is based on the source-filter-model of speech. The digital filter the coefficients are assessed automatically from a frame of natural speech. There are two important parts in LPC 1. Current speech sample of $y\ (n)$ and estimated or predicted from a finite number of previous $p$ samples $y\ (n\text{-}1)$ to $y(n\text{-}k)$ by a linear combination with small error term $e(n)$ called residual signal. Thus,

$$y(n) = e(n) + \sum_{k-1}^{\rho} a(k)y(n-k)$$

and

$$e(n) = y(n) - \sum_{k-1}^{\rho} a(k)y(n-k) = y(n) - \ \bar{y}(n)$$

where $y(n)$is a predicted value, $p$ is the linear predictor order, and $a(k)$ are the linear prediction coefficients which are found by minimizing the sum of the squared errors over a frame. In synthesis generate the pulse excitation by a train of impulses for voiced sounds and by random noise for unvoiced. The excitation signal is then extended and filtered with a digital filter for which the coefficients are $a\ (k)$. The filter order is typically between 10 and 12 at 8 kHz sampling rate, but for higher quality at 22 kHz sampling rate, the order needed is between 20 and 24 (Klein et al. 1998, Karjalainen et al. 1998). The coefficients are usually updated every 10-15 ms because of length of vocal cord which of 17 cm.

### Sinusoidal Models

In this method Sinusoidal models are built on a well-known hypothesis that the speech signal can be represented as a sum of sine waves with time-varying amplitudes and frequencies (McAulay et al. 1986, Macon 1996, Kleijn et al. 1998). In the basic model, the speech signal $s(n)$ is modeled as the sum of a small number $L$ of sinusoids

$$s(n) = \sum_{l-1}^{L} A_i \cos\,(w_i n + \emptyset_1)$$

where $A_l(n)$ and f $_l(n)$ represent the amplitude and phase of each sinusoidal component associated with the frequency track w $_l$. To find these parameters $A_l(n)$ and f $_l(n)$, the DFT of windowed signal frames is calculated, and the peaks of the spectral magnitude are selected from each frame (see Figure 2). The basic model is also known as the McAulay/Quatieri Model. The basic model has also some modifications such as ABS/OLA (Analysis by Synthesis / Overlap Add) and Hybrid / Sinusoidal Noise models (Macon 1996)[12].

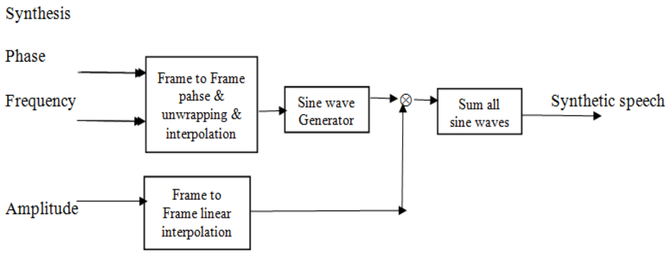

**Fig. 2. Block diagram of Sinusoidal analysis / synthesis system (Macon 1996)**

While the sinusoidal prototypes are perhaps very appropriate for representing periodic signals, such as vowels and voiced consonants, the representation of unvoiced speech becomes problematic (Macon 1996).

### High-Level Synthesis

One of the model of speech synthesis is high-level synthesis the input text is capable to produce the acoustic output in the form of speech. A proper implementation of this is the fundamental challenge in all present systems and will probably be for years to come. The model consists of three main stage.

1. In this high level synthesis the process start fron text preprocessing in this numerals, special characters, abbreviations, and acronyms are expanded into full words.
2. The deep pronunciation analysis where done for specfic words, including homographs and proper names, are determined.
3. Final stage of the model was prosodic analysis where the prosodic features of speech are determined.

After this process, the information is given to drive some low-level system. The type of used data depends on the driven system. For example, some of the parameter of speech was formant synthesizer, fundamental frequency, formant frequencies, duration, and amplitude of each sound segment is needed.

- One of the features speech is voice quality of the speech which contains largely constant voice characteristics over the spoken utterance, such as loudness and breathiness. For example, 1.The angry voice gives the breathy, loud, and has a tense articulation with abrupt changes and on the other hand sad voice is very quiet with a decreased articulation precision.
- Another features of speech is Pitch contour and its dynamic changes carry important emotional information, both in the general form for the whole sentence and in small fluctuations at word and phonemic levels. The most important pitch features are the general level, the dynamic range, changes in overall shape, content words, stressed phonemes, emphatic stress, and clause boundaries.
- The parameter of speech is time characteristics contain the general rhythm, speech rate, the lengthening and shortening of the stressed syllables, the length of content words, and the duration and placing of pauses.

## Evaluation of speech quality and Test

After the generation of synthetic speech the quality of the speech measure using parameters intelligibility, naturalness, and suitability for used application (Klatt 1987, Mariniak 1993). In some applications, for example reading machines for the blind, the speech intelligibility with high speech rate is usually more important feature than the naturalness. The evaluation methods are usually proposed to test speech quality in general, but most of them are suitable also for synthetic speech. It is very difficult, almost impossible, to say which test method provides the correct data. In a text-to-speech system not only the acoustic characteristics are important, but also text pre-processing and linguistic realization determine the final speech quality.

## Segmental Evaluation Methods

In whole speech signal we first divide the speech in the frame that frame is called as segment. With segmental evaluation methods only a single segment or phoneme intelligibility is tested. The very commonly used method to test the intelligibility of synthetic speech is the use of so called rhyme tests and nonsense words. The rhyme tests have several advantages (Jekosh 1993). The number of stimuli is reduced and the test procedure is not time consuming. Also naive listeners can participate without having to be trained and reliable results can be obtained with relatively small subject groups, which is usually from 10 to 20.

## Nonsense words and Vowel-Consonant transitions

In most of method the use syllable, that syllable consist of nonsense words (logotoms), mostly transitions between vowels (V) and consonant (C) is one of the most commonly used evaluation method for synthetic speech. This method gives the high error rates and excellent diagnostic material especially when open response set is used. Usually a list of VC, CV, VCV or CVC words is used, but longer words, such as CVVC, VCCV, or CCCVCCC, are sometimes needed. Especially when testing diphone-based systems, longer units must be used to test all CV-, VC-, VV-, and CC-diphone-units.

## Sentence Level Tests

Several sets of sentences have been developed to evaluate the comprehension of synthetic speech. Sentences are usually chosen to model the occurrence frequency of words in each particular language. Unlike in segmental tests, some items may be missed and the given answer may still be correct, especially if meaningful sentences are used (Pisoni et al. 1980, Allen et al. 1987).

## Mean Opinion Score (MOS)

This method used to evaluate speech quality. It is also appropriate for overall evaluation of synthetic speech. The test of speech evaluation, the MOS is consist of five level scale from bad (1) to excellent (5) and it is also known as ACR (Absolute Category Rating). DMOS is an impairment grading scale to measure how the different disturbances in speech signal are perceived. Overall tests together provides lots of useful information, but is on the other hand very time-consuming. The test methods must be chosen carefully because there is no sense to have the same results from two tests. It is also important to consider in advance what kind of data is needed and why. It may be even reasonable to test the method itself with a very small listening group to make sure the method is reasonable and will provide desirable results. Finally the objective of this test is to develop as well as speech synthesizers. Feedback from real users is needed and necessary to develop speech synthesis and the assessment methods.

## Conclusion

In this paper, the final output of text to speech is synthesized speech can be evaluated by many methods and at several levels. All this methods give best information on speech quality, but it is easy to see that there is no test to give the one and only correct data there was variation for different database. Perhaps the most suitable way to test a speech synthesizer is to select several methods to assess each feature separately. For example using segmental, sentence level, prosody, and overall tests together provides lots of useful information, but is on the other hand very time-consuming.

## Acknowledgment

The author P.M. Ghate is working as an Assistant Professor at JSPM’s Rajarshi Shahu College of Engineering, Pune, India. Also we would like to thank the institute.

## REFERENCES

Abadjieva, E., Murray, I. and Arnott, J. 1993. Applying Analysis of Human Emotion Speech to Enhance Synthetic Speech. *Proceedings of Eurospeech 93* (2): 909-912.

Belhoula, K. 1993. Rule-Based Grapheme-to-Phoneme Conversion of Names. *Proceedings of Eurospeech 93* (2): 881-884.

Benoy Kumar Thakur, Bhusan Chettri and Krishna Bikram Shah**,** 2012. **"**Current Trends, Frameworks and Techniques Used in Speech Synthesis – A Survey" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-2, May 2012.

Carlson, R., Granström, B. and Nord, L. 1990. Evaluation and Development of the KTH Text-to-Speech System on the Segmental Level. *Proceedings of ICASSP 90* (1): 317-320.

Donovan, R. 1996. *Trainable Speech Synthesis*. PhD. Thesis. Cambridge University Engineering Department, England.<ftp://svrftp.eng.cam.ac.uk/pub/reports/donovan_thesis.ps.Z.

Dutoit, T. 1997. An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers, Dordrecht, 1997.

Flanagan, J. L. 1972. Speech analysis, synthesis and perception, Springer-Verlag.

Klatt, D.H. 1980. "Software for a cascade/parallel formant synthesizer," J.Acoust. Soc. Amer., vol. 67, pp. 971–995.

Kortekaas, R. and Kohlrausch, A. 1997. Psychoacoustical Evaluation of the Pitch-Synchronous Overlap-and-Add Speech-Waveform Manipulation Technique Using Single-Formant *he Acoustical Society of America, JASA*, Vol. 101

Logan, J., Greene, B. and Pisoni, D. 1989. Segmental Intelligibility of Synthetic Speech Produced by Rule. *Journal of the Acoustical Society of America, JASA* vol. 86 (2): 566-581.

Macon, M. 1996. *Speech Synthesis Based on Sinusoidal Modeling*. Doctorial Thesis, Georgia Institute of Technology

Madhavi R. Repe, Shirbahadurkar, S.D. and Smita Desai, 2010. "Prosody Model for Marathi Language TTS Synthesis with Unit Search and Selection Speech Database**"** International Conference on Recent Trends in Information, Telecommunication and Computing

Quartier, T. 2002. Discrete-time speech signal processing. Prentice-Hall, 2002.

Rohit Kumar, S. P. Kishore, 2004. "Automatic Pruning of Unit Selection Speech Databases for Synthesis without Loss of Naturalness", International Conference on Spoken Language Processing (Interspeech - ICSLP), October 2004, Jeju Korea

Sangramsing Kayte and Bharti Gawali, 2015. Article: A Text-To-Speech Synthesis for Marathi Language using Festival and Festvox. *International Journal of Computer Applications* 132(3):35-41, December 2015. Published by Foundation of Computer Science (FCS), NY, US

Shirbahadurkar, S.D. and Bormane, D.S. 2009. "Marathi Language Speech synthesizer using concatenative synthesis strategy" Machine Vision, ICMV '09. Second International Conference.

*******