IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11575

Marathi Text to Speech Conversion Using Concatenative Approach

Patekar Komal[1], Shivani Pardeshi[2], Pratik Watane[3], Atharva Thosar[4], and K.P.Birla[5]

[1,2,3,4]Student Department of Computer Engineering, K.K.Wagh Institue of Engineering Education and Research.

[5]Research Guide Department of Computer Engineering, K.K.Wagh Institue of Engineering Education and Research.

Abstract—Language resources are essential to the developmentof text-to-speech (TTS) systems. The goal of TTS technology Text-to-Speech conversion is no longer to simply make machines talk, but to make them sound like people of different ages and genders.The quality of TTS systems synthesizers is evaluated from avariety of perspectives, including intelligibility, naturalness, and preference of the synthesized speech, as well as human perception factors, such as comprehensibility. TTS using concatenative TTS relies on high-quality audio clips, which are then combined to form the speech. At the first step voice used for searching is recorded manually and stored in system for further checking andconversion from a range of speech units. The transformation is done from whole sentences to syllables that are further labeled and segmented by linguistic units from phones to phrases and sentences forming a huge database. During speech synthesis, a Text-to-Speech engine searches such database for speech units that match the input text, concatenates them together and produces an audio file in the same directory , which containsthe final output.

Index Terms—Concatenative speech synthesis, Unit Size, Syl- labification,Spectral Noise Reduction, Satistical Parametric Syn- thesis model.

I. INTRODUCTION

We human beings use speech as the most basic form of daily communication.Speech is the primary means of communication between people. Speech synthesis is an automatic generation of speech waveforms. The dream of producing a talking machine started at the 18th century has made a progress. With the increase in advancements of technology speech synthesis also has increased.This advancements has increased the degree of synthesis but the problem of intelligibility and naturalness still remains to be major roadblock in speech synthesis systems. Ancient and medieval Maharashtra included the empires of the Satavahana dynasty, Rashtrakuta dynasty, Western Chalukyas, Mughals and Marathas it is bordered by the Arabian Sea to the west and the Indian states of Karnataka, Telangana, Goa, Gujarat, Chhattisgarh, MadhyaPradesh and the Union territory of Dadra and Nagar Haveli. The major rivers of the state are Godavari, Krishna, Narmada and Tapi. The state has several tourist destinations including the popular Hindu places of pilgrimage, Pandharpur, Dehu and Alandi. Places with wide appeal include Hazur Sahib Nanded at Nanded, and Saibaba shrine at Shirdi. Maharashtra is the most urbanized state in India, with large cities besides the capital Mumbai such as Pune, Nagpur, Nashik and Aurangabad . So, it's an important issue to build Marathi TTS which is reliable, intelligent and user friendly system to give thosepeople a chance to use the technologies like text messages, emails, and web sites using their native language. Also thisis at the ancient times nowadays people are spread over huge areas so its scope is not only limited to western india but also various parts of the world.

This fact makes speech synthesis in Marathi Language an important field for investigation and improvement for the major languages including Marathi.Speech synthesis is a challenging task as the text that is input may have ambigious form .This technique is based on concatenative approach of speech units that required a prosodic modification algorithm to adjust the prosodic features of the stored speech units to the desired output values.Also Speech synthesizer or as it known Text-to-Speech system While communicating with the computer systems, interacting with the help of writtentext and images make the use of computers difficult for visually , physically impaired and illiterate masses.Also dueto increase in requirement of the speed at which data is entered in the system needs to be increased.The technique of Conventional one to one word typing needs to be improved to meet the demands. To challenge this method one new method is proposed which involves writing the words in user's own language which the system the convert into Marathi language and produce output.

There are many paid services for Marathi Text to Speech Conversion such as:Voicemaker ,Indiantts ,Play.ht Although these services provide good output, they charge a lot of money for Text-to-Speech conversion. The ones that are free do

DOI: 10.17148/IJARCCE.2022.11575

not sound native and natural. Some services such as TTSFree.com claim that they provide a free service but are actually paid services. Thus providing a completely free Marathi Text to Speech service is our motivation.

The purpose of this project is to achieve the ease with which marathi text to speech output can be understood as well as how to increase the naturalness i.e the resemblance of input text to orignal text in Marathi Language. Most text to speech output lacks both inteligibility and naturalness.It is challenging to achieve naturalness for Marathi Text to Speech Conversion because there are many paid services available and the free ones sound unnatural. Hence, we try to achieve naturalness for Marathi Text to Speech Conversion.In the first step, voice actors record a range of speech units, from whole sentences to individual syllables that are then labelled and segmentedby linguistic units, from phones to phrases and sentences,resulting in a huge database.A Text-to-Speech engine searchessuch a database for speech units whose input matches the inputtext, concatenates them and creates an audio file.

In theory articulatory synthesis attempts to model the human speech directly, but its implementation is extremely difficult.Therefore a hybrid approach is selected which comprises of concatenative speech synthesis technique. It is not as efficient as articulatory method but it is feasible and easy to implement. As goes with the saying every technique has its own advantages and disadvantages.

Concantenative speech synthesis (CSS), also known as unit selection speech synthesis, is one of the two primary modern speech synthesis techniques together with statistical paramet- ric speech synthesis. It uses prerecorded speech samples to concatenate as well as match tokenized words for final con- version.The resulting output of the system is high naturalness of the produced speech as it used precorded samples.

A. Marathi text to speech converter system

Syllabic based concatenative speech synthesis is used for reducing the size of the database. This method creates new words using existing words and syllables.Syllabification is the task of detecting syllable boundaries within words. The boundaries depend on the phonetic structure of thosewords and on phonotactic constraints of the respective language. Syllabification technique include method which consists of neural networks.Further a classification neural network is used in the technique to improve the results of syllabification.Syllabification that uses position as major parameter was used in the technique. Objective spectral noise reduction technique was also used to enhance the working of algorithm.Objective spectral estimation is used to eliminate the audio distortion present in concatenated samples. Majorly it eliminates the spectral distortion.After sufficient elimination of distortion and noises the newly formed words are added back to database.Finally Satistical Parametric Synthesis modelis used for text to speech conversion.

1) Concatenative Speech Synthesis: Concantenative speech synthesis (CSS), also known as unit selection speech synthesis, is one of the two primary modern speech synthesis techniques together with statistical paramet- ric speech synthesis. It uses prerecorded speech samples to concatenate as well as match tokenized words for final con- version.The resulting output of the system is high naturalness of the produced speech as it used precorded samples.

This Speech Synthesis is a technique which uses short recorded samples of speech to synthesize words or speech of larger length i.e complete words or sentences.This unit size samples may vary in length from upto 10 miliseconds to 1 seconds.It is used to generate user specified sequences from the database of pre-recorded samples that is created using unit length samples.Syllabification: The given input text was segmented intoa series of small texts that are known as sequence syllabic speech units.Speech in indian language is diverse in nature and contains different choice of syllable. Each syllable contains a different choice of phoneme.The syllable unit is used to improve the results of output.There is still a major debate among many linguist regarding the exact defination of syllableand its usage. According to the Sonority Theory of a sound, syllabe is related to the characteristics of the sound sample which it contains.

2) Spectral Noise Reduction: Noise Reduction is a spectral noise gate designed for removing unwanted noise from a variety of audio sources. By targeting specific frequency ranges, its spectral gate effectively filters out noise with a fair degree of customizability over attack and release times. Specs and useful features spectral subtraction is used in this research as a method to remove noise from noisy speech signals inthe frequency domain. This method consists of computing the spectrum of the noisy speech using the Fast Fourier Transform (FFT) and subtracting the average magnitude of the noise spectrum from the noisy speech spectrum. This Spectra Layers Elements noise reduction tutorial looks at how to reduce noisein a song, in a way that is much more effective than usinga noise reduction plugin. The way that most noise reduction software takes a full spectrum snapshot in order to work meansthat it also takes part of the audio spectrum that you want to keep. In Spectra Layers, you can remove the hissing sound from analogue tape much

DOI: 10.17148/IJARCCE.2022.11575

more efficiently, keeping all the parts of the spectrum that you do not want to lose. It’s a much more efficient way of reducing noise with software.

3) Satistical Parametric synthesis model.: Satistical Para- metric Synthesis model: When we talk about a model-based approach to speech synthesis, particularly when we wish to learn this model from data, we generally mean a statisti- cal parametric model. The model is parametric because itdescribes the speech using parameters, rather than stored exemplars. It is statistical because it describes those param- eters using statistics (e.g., means and variances of probability density functions) which capture the distribution of parameter values found in the training data. The remainder of this article will focus on this method for speech. Synthesis statistical parametric speech synthesis uses the source-filter representa- tion of speech, the spectrum, excitation, and duration can be controlled and modified separately.

Historically, the starting point for statistical parametric speech synthesis was the success of the HMM for automatic speech recognition. No-one would claim that the HMM is a true model of speech. But the availability of effective and efficient learning algorithms (Expectation– Maximization), automatic methods for model complexity control (parameter tying) and computationally efficient search algorithms (Viterbi search) make the HMM a powerful model. The performance of the model, which in speech recognition is measured using word error rates and in speech synthesis by listening tests, depends critically on choosing an appropriate configuration. The two most important aspects of this configuration are the parameterization of the speech signal (the ‘observations’ ofthe model, in HMM terminology) and the choice of modelling unit. Since the modelling unit is typically a context-dependent phoneme, this choice means selecting which contextual factorsneed to be taken into account.[7]

B. Format of input text

The Indian Language Marathi in the digital computer in the form of ISCII (Indian Script Code for Information Interchange), transliteration scheme of various fonts and UNI- CODE.The Indian language have common phonetic base.The scheme of ISCII has largely been overtaken by UNICODE. The input can hence be given to the system with the help of various Devanagari keyboard layouts.

C. Speech Generation Component

The synthesis of input text is and its conversion into speech is commonly taken care by speech generation component . The available recorded samples for each text units is used to generate speech. All the possibility of such text is stored in a the database in memory.This samples is further used for concatenation of words to convert text to speech. As there is availability of cheap memory and computation power the storage of large number of units has been possible andalso their retrieval in real time is feasible. Further, using an inventory of speech units is referred to as unit selection approach however, it can also be referred to as data-driven approach or example based approach for speech synthesis.

II. REVIEW OF PAPERS

This blog discusses Satistical parametric synthesis used in this model.This model uses utilizes recorded human samples.We use a set of function and set of parameters to modify the voice.In this technique, we generally have two parts. The training and the synthesis. During training, we extract a setof parameters that characterize the audio sample such as the parameter that describe the text. We then try to estimate those parameters using a statistical model. The one that has been proven to provide the best results historically is the Hidden Markov Model (HMM).[1]

The open source Festival TTS Engine was used to im-plement Marathi text to speech conversion for Marathi Lan- guage. This system is developed using diphone concatenation approach in its waveform generation phase. NLP modules include text processing methods and text conversion methods that were used to process and convert the given input textinto a series of acceptable tokens that will be used by the model.The technique of syllabification is derived from this table. Festival’s default syllabification algorithm based on sonority sequencing principle is used to syllabify the Marathi words. Besides the default syllabification algorithm, our lexi- con has also been syllabified along with pronunciation. [2]

This paper discusses a concatenative approach for creating voice naturalness by reducing the database size. The scope of this paper is limited to concatenative analysis. Articulatory and Parametric analysis are not discussed.[3]

This Paper focuses on the development of expressive Text- to-Speech synthesis techniques for Marathi (spoken in Maharashtra, India) language.The Pitch contour is one of the important properties of speech that is affected by emotional speech.[4]

This paper investigates the method for ”Devanagari” Script to Speech Conversion” as applied to Marathi language by developing the appropriate software. Neural networks are primarily used which does the work of pattern matching and character recognition. Search and match technique has been primarily.[5]

III. METHODOLOGY

A. User feedback or input

We take input from the user through the keyboard.Thisstring is further processed for conversion, Firstly the inputis tokenized into a sequence of words. If the phonemes are available for current word, then we simply proceed forward for conversion. Else if not available, a file is created which contains the new phonemes and a index number is provided tothe file. After this conversion it is added to the program file.Incase of correct input as well as the phonemes availability, the phonemes from the sequence are matched with the database. When the matching phoneme is found its index is returned from the database folder and finally the file corresponding to that phoneme is selected, and is added into the pronunciation file. This file will be later played which contains the final converted speech. Create a data-set which contains maximum number of phoneme files.Adding new words to the pronunci- ation file. To increase the quality of output the old recording’s has to be replaced with new recording’s which are good in quality.Better tokenization algorithm as to be applied. Diverselibrary of Marathi file’s has to be made.

B. Algorithm

1) Step 1: Start the program.

2) Step 2: Input Marathi text ’P’ in the text box ’T’.

3) Step 3: : Check if there is any input ’P’ in the text box ’T’ if There is any input ’P’ present in the text box ’T’, thengo to step ’4’. otherwise End

4) Step 4: :Tokenize the text ’P’ into words w1, w2, w3. wn and add to a list ’l’.

5) Step 5: Check if the word ’w’ is present in the pronun- ciation file if The word ’w’ is present in the pronunciation file ’Pf’, then get the audio file name sequence ’s’.otherwise Tokenize the word ’w’ into its corresponding phonemes p1, p2, p3. pn and create a sequence ’s’ using the corresponding
audio file names pa1, pa2, pa3. pan.

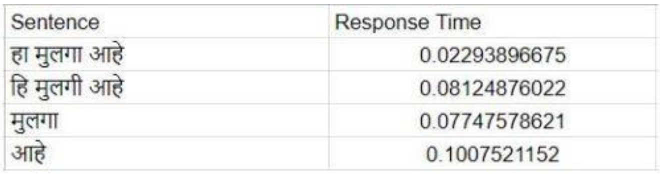
6) Step 6: Add the sequence ’s’ to the result list ’r’.

7) Step 7: Check if there are any words in the list ’l’ if There is at least one word present in the list, then goto step ’5’.otherwise Using the sequence ’s1’ from the result list ’r’, merge the corresponding audio files a1, a2...... an from the database ’d’ together to form the resultant audio file ’A’ forthe input text ’P’.

8) Step 8: The output ’A’ is provided as a .wav file.

9) Step 9: End

C.



Sentence	Response Time
हा मुलगा आहे	0.02293896675
हि मुलगी आहे	0.08124876022
मुलगा	0.07747578621
आहे	0.1007521152

DOI: 10.17148/IJARCCE.2022.11575

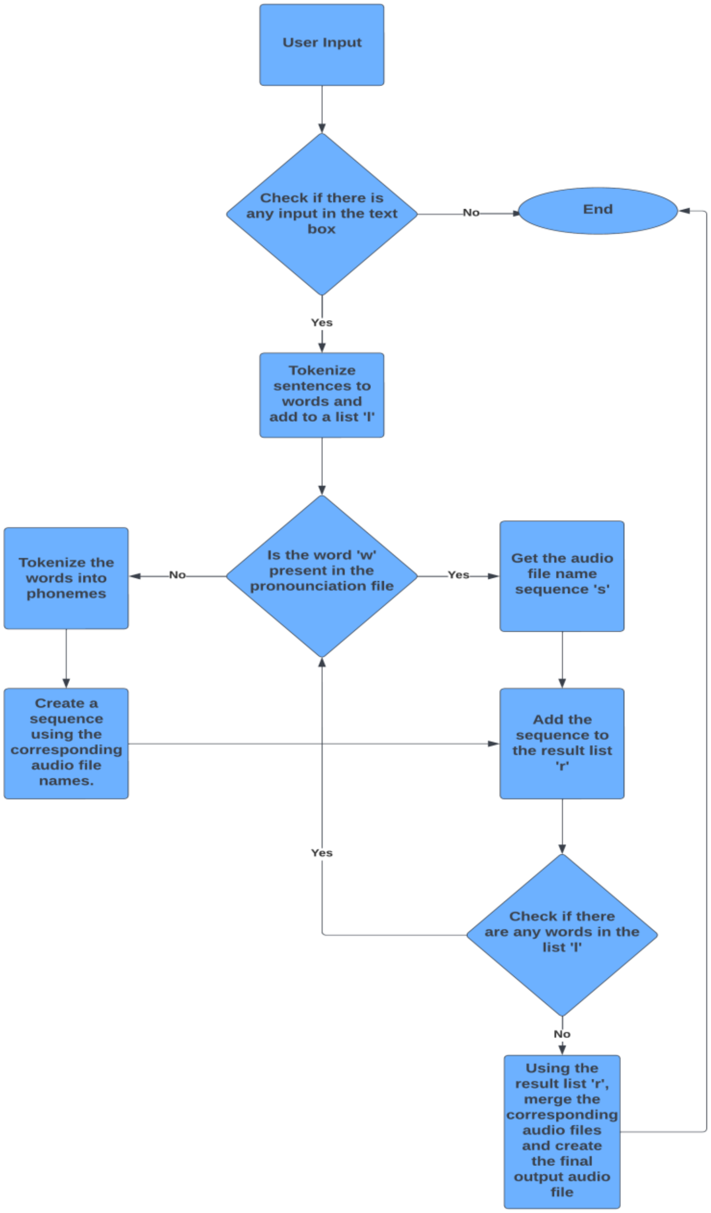

User Input
Check if there is any input in the text box
No
End
Yes
Tokenize sentences to words and add to a list 'l'
Is the word 'w' present in the pronounciation file
No
Tokenize the words into phonemes
Yes
Get the audio file name sequence 's'
Create a sequence using the corresponding audio file names.
Add the sequence to the result list 'r'
Yes
Check if there are any words in the list 'l'
No
Using the result list 'r', merge the corresponding audio files and create the final output audio file

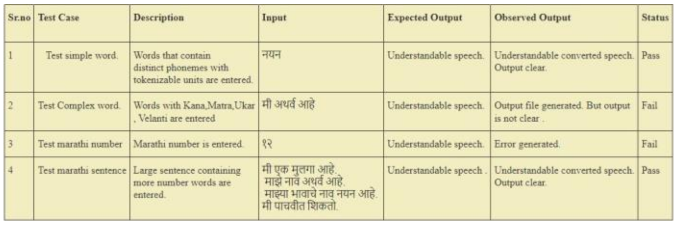
Fig. 1. Flowchart

D. Efficiency Issues

No. of opinions recorded = 12 Mean opinion score = 3The mean opinion score '3' shows that although the speech was understandable to a certain extent, it was not natural.The samples were limited which also contributed to the problem. Also the quality of device used to record the samples was improved to increase the naturalness. But it was observed a better algorithm might have been there if we were to increase the naturalness.

V. RESULT

Test Case



Sr.no	Test Case	Description	Input	Expected Output	Observed Output	Status
1	Test simple word.	Words that contain distinct phonemes with tokenizable units are entered.	नयन	Understandable speech.	Understandable converted speech. Output clear.	Pass
2	Test Complex word.	Words with Kana,Matra,Ukar , Velanti are entered	मी अथर्व आहे	Understandable speech.	Output file generated. But output is not clear .	Fail
3	Test marathi number	Marathi number is entered.	१२	Understandable speech.	Error generated.	Fail
4	Test marathi sentence	Large sentence containing more number words are entered.	मी एक मुलगा आहे. माझे नाव अथर्व आहे. माझ्या भावाचे नाव नयन आहे. मी पाचवीत शिकतो.	Understandable speech .	Understandable converted speech. Output clear.	Pass

Fig. 2. Test case

VI. CONCLUSION

In this paper we were supposed to achieve naturalness to theoutput using concatenative approach, but our hypothesis failed as we were unable to provide the output with naturalness. In this presented paper the naturalness is not achieved as per expectations. We have successfully created a locally hosted Marathi Text-to-speech web service with less naturalness

IV. EXPERIMENTAL SETUP

A. Data set

A user created data set will be used for Marathi speech synthesis.Attributes are audio and transcripts.A database of 48audio samples were created and were used for the process of identification of words.

B. Performance Parameters

VII. FUTURE WORK

In the future, we will make this service available onlineand also improve the service by including the user feedback feature, where the quality of the application is improved by taking feedback from the user on how correct the output is.

ACKNOWLEDGMENT

It gives us great pleasure in presenting the paper on ’MarathiText to Speech Converter’.We express deep and sincere gratitude to **Prof. K.P. Birla** whose guidance, encouragement, suggestion and very constructive criticism have contributed immensely to the evolution of ideas on our project . We are extremely grateful for the generous assistance. We are also grateful to **Prof. Dr. S. S. Sane**, Head of Computer Engineering Department, K. K. Wagh Institute of Engineering Education and Research, Nashik for his indispensable support, invaluable assistance and suggestions. We would like to thank the staff of K. K. Wagh Institute of Engineering Education and Research for their support and assistance. A special thanks to our friends and colleagues for giving valuable feedback and suggestions that helped us improve

REFERENCES

[1] Sergios Karagiannakos, A review of the best text to speech architectures with Deep Learning 2021.

[2] Sangramsingh Kayte, Bharti Gawali A Text-To-Speech Synthesis for Marathi Language using Festival and Festvox. 2015.

[3] Smita P. Kawachale, J.S.Chitode Position based syllabification and ob- jective spectral analysis in Marathi text to speech for naturalness 2014.

[4] Manjare Chandraprabha Anil, S.T.Shirbahadurkar Speech Modification for Prosody conversion in expressive Marathi Text-To-Speech 2014.

[5] S.D. Shirbahadurkar D.S. Bormane, Marathi Language Speech Synthe- sizer Using Concatenative Synthesis Strategy, 2009.

[6] Mrs.Madhuri Tasgonkar, Prof. C.V.Joshi, Prof. N.B.Pasalkar Script to speech conversion for Marathi Language. 2003.

[7] Shaikh Shadab Shakil , Manjare Chandrakant Patil Cognitive Devanagari (Marathi) Text-to-Speech SystemCognitive Devanagari (Marathi) Text-to- Speech System