International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

Text to Speech Synthesis using Fraction Based Waveform Concatenation and Optimal Coupling Smoothing Technique

Saranya S, A.Rathinavelu, Jayashree C

Abstract: ***Text to Speech System is a Speech Synthesis application that converts a text to speech. The current project focuses on developing a TTS System for the Tamil Language with the Synthesis Technique as Unit Selection Synthesis. Letter Level Segmentation of an input text helps in the reduction of corpus size compared to Syllable Level Segmentation. The segmented units are retrieved with respect to Unicode values, concatenated and the synthesized speech is produced. Intelligibility and Naturalness of the spoken word can be improved using the Smoothing Techniques. Optimal Coupling Smoothing Technique is implemented for the smooth transition in between the concatenated speech segments to create continuous Speech output like human voice. Fraction based Waveform Concatenation method is used to produce the intelligible speech segments as output from the pre-recorded speech database.***

Index Terms :Text to Speech Synthesis, Letter level Segmentation, Smoothing, Fraction based Waveform Concatenation, Optimal coupling Technique

I. INTRODUCTION

Text to Speech (TTS) is a Speech Synthesis technique that is used to create a Spoken Sound of text from the digital documents. It enables reading of digital information for the visually disabled person, or used to augment the reading of text information. speech synthesis is artificial speech producing. Any person who is not able to speak but can use an interface has the potential to use a text-to-speech system to provide themselves with a voice. Synthesized speech can be generated by concatenation of sections of recorded speech segmrnts that are stored in the database. The quality of speech synthesis is based on the Naturalness and Intelligibility obtained. Fig 1 illustrates the generation of synthetic speech. The task of proposed speech synthesis for the Tamil language includes Pre-processing of both the Input text and the Speech Corpus. Segmentation is Letter Level and the corresponding speech units are selected based on the Unit Selection Synthesis technique using python library[3].

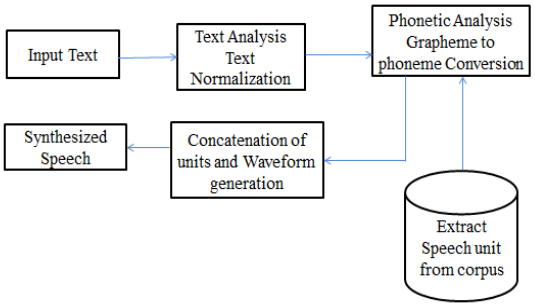

Fig. 1. Speech Synthesis System

Letter level segmentation of input text helps in the reduction of corpus size compared to Syllable Level Segmentation[10]. Smoothing applied in the proposed TTS system enables the smooth transition between concatenated speech segments by removing the discontinuities between speech segments. Prosody, text pre-processing and pronunciation need to be embedded to produce pleasant speech. The improvement in voice quality and language has been analysed.

II. LITERATURE REVIEW

There are several techniques available for developing Text-To-Speech synthesis system. Those techniques generally falls under three major categories as follows

A. Formant Synthesis

In Formant Synthesis, the vocal tract transfer function is built by simulating the frequencies. The synthesis is done and the speech is created based on the models of the human speech organ. The acoustic tube model is used by the formant synthesizer for generating the sound from a source. It is periodic for voiced portion of sounds and noise portion for obstruent sounds. This source signal is placed into the vocal-tract model.

Manuscript received on April 02, 2020.
Revised Manuscript received on April 15, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author
Saranya S, Department of Computer Science and Engineering, Dr Mahalingam College of Engineering and Technology, Pollachi, India. Email: saran41sakthi@gmail.com
Dr.A.Rathinavelu, Department of Computer Science and Engineering, Dr Mahalingam College of Engineering and Technology, Pollachi, India. Email:
starvee@yahoo.com
Jayashree C, , Department of Computer Science and Engineering, Dr Mahalingam College of Engineering and Technology, Pollachi, India. Email: shree.rcj@gmail.com

This produced signal passes via the oral and nasal cavity, finally it penetrates through a radiation component, which simulates the load propagation characteristics and speech pressure waveform is produced. The human speech samples are not used in the formant synthesis but depends on the rules written by linguists to generate the parameters permitted by the speech synthesis, to deal with the movement from one phoneme to another is the co-articulation of speech[11].

Linguists have analysed spectrograms and derived the rules of stages of evolution of frequency change is called formants. The optimal rule to do the synthesis is not yet known. Also the speech waveform production is a complex process that, rules can only model the features of the speech waveform.

B. Articulatory Synthesis

Articulatory synthesis defined as the technique synthesis of speech depends on the vocal tract models combined with the articulation processes of human voice. Articulatory speech synthesis outputs the natural speech produced is as accurate with naturalness. A synthetic model of human physiology is created and making it speak. Articulatory synthesis systems encloses with components of :

A module for the creating the vocal tract movements, and the other module for converting this movement information into a continuous sequence of vocal tract geometries, and final module for creation of acoustic signals on the basis of articulatory information. This synthesis creates speech by direct modelling of the articulatory behaviour of human, this is the widely used method to produce quality speech[11].

C. Concatenative Speech Synthesis

Concatenative speech synthesis method includes producing the synthesized speech by merging pre-recorded units of speech by phonemes, di-phones, syllables, words or sentences. This synthesis involves the selection of appropriate units from the speech database and join selected units and some signal processing, smoothing the boundary points are to be done. Here, speech is produced by concatenating matched audio units from database which consists of different sizes of speech units[11]. Concatenative speech synthesis includes different methods such as Unit Selection synthesis, Di-phone synthesis etc.

III. EXISTING FRAMEWORK

Text-to-speech (TTS) systems which uses the concatenation technique produces continuous speech by fetching waveform units from speech databases. Corpus contains Speech units obtained through letter level segmentation. Unit Selection synthesis is used for synthesis process. Since it uses pre-recorded human voice, naturalness is achieved in speech output [3]. More number of systems uses databases with a large number of available segments with varied characteristics. The large-database synthesis methods mostly concentrate on segment choice and search algorithms. It retrieves the matched speech units from the desired text. Concatenative speech synthesis with a few hundred to thousand units of speech waveforms for every speaker will have more discontinuities at segment boundaries. Each speech segment should be changed to fit the appropriate properties.

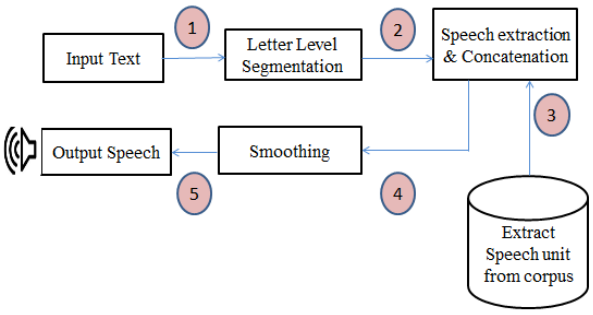

Fig. 2. Block Diagram of Existing System

The spectral characteristics of the beginning and ending of each sound should be moulded[7] for smoothening the transition between interconnecting units. It is also needed to have the knowledge about when the spectral smoothing is needed.

IV. PROPOSED WORK

Corpus has speech units created by letter level segmentation. The appropriate speech units are retrieved from the speech database using unit selection synthesis technique. Since it uses pre-recorded speech, naturalness is achieved in speech output. But the intelligibility of the system and Co-articulation between consecutive words is of greater problem. Unit Selection uses large database but it concentrate on segment choice and search algorithms. The corpus has enough sample units for a close match for each phoneme. Fraction based Waveform Concatenation technique to produce quality speech segments from the speech database. For enhancing the naturalness of the system Smoothing Technique can be used after Concatenation synthesis technique for removing discontinuity between two segments [5]. Fraction based Waveform Concatenation technique implemented to discontinuities from voice database. To improve the naturalness of the method Smoothing Technique can be used to eliminate discontinuity between two segments after Concatenation synthesis technique. Optimal Coupling smoothing technique is to improve the naturalness of speech.

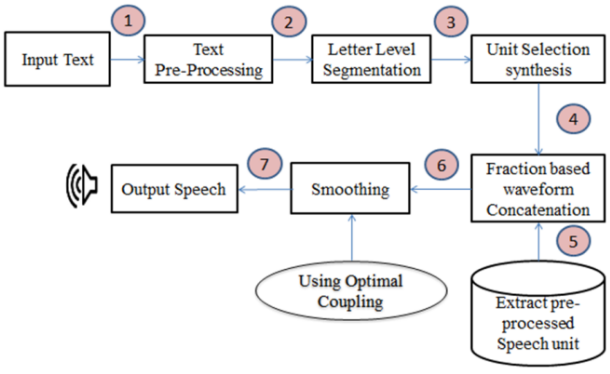

Fig.3. Block Diagram of Proposed System

A. *Steps Involved In Proposed System*

- Text is given as input to the system in English language and converted to Tamil text
- Pre-processing of the input text and the words in Speech Corpus is done.
- Input text Segmentation is done in letter level
- Units are selected and concatenated by unit level synthesis and Fraction based concatenation.
- Optimal Coupling Smoothing is done between speech segments to produce the output to achieve naturalness.

V. METHODOLOGY

The Proposed TTS system mainly has four steps as to follows. In the proposed system, the input is given as "kuruvi". In second step, the text preprocessing is done. In third step, the word is split into letters and transliterated into tamil letters by unicode conversion. If the input text is "kuruvi", it is converted to "குருவி" and splitted as கு(ku), ரு(ru), வி(vi). The third component will be generating the speech includes selecting the audio segments from the speech corpus, and then concatenating them as a single audio file by using fraction based concatenation[1].

A. *Speech Corpus Development*

The speech corpus development for the tamil language is difficult than the English language. Audio corpus is developed that contains the audio file for each letter in Tamil Whole Corpus Collection is done by individual Speaker. The audio for each and every letters are not recorded separately, instead a word that incorporates the corresponding letter is recorded as speech waveform and the voiced portion that matches to the particular letter is chopped and stored as a new audio file.

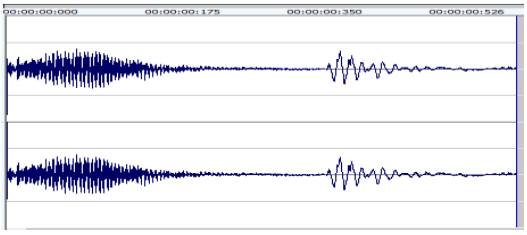

Fig. 4. waveform of the letter 'அ'

But after chopping, it becomes so hard to extract the audios of certain letters exactly, because the consonant letters will often be co-articulated with their prefixes[9]. In such cases, it is required to separate the portion of the waveform belonging to the consonant exactly. It might not be meaningful at first, but while concatenating, it exhibits its nature of co-articulation along with its neighbour phonemes and thus helps in reproducing the actual sound back. For this fraction based concatenation is used for concatenating speech units[3].

Table- I: Speech Corpus

	அ	ஆ	இ	ஈ	உ
க்	கண் கர்வம்	காசு காந்தம்	கிளி காகிதம்	கீதம் கீர்த்தி	குருவி குணம் அடக்குமுறை அங்குலம்
ச்	சக்கரை சங்கமம்	சாட்டை சாதி	சிறிய சின்ன	சீட்டு	சுட்டு சுடு

B. *Pre-Processing Of The Text*

In TTS system, pre-processing is the first process has to be done for the text. The special symbols and the punctuation marks are identified and those are eliminated to produce the actual text. This process is called Text Normalization. For e.g., the input text is "kuruvi?" then after pre-processing the output will be produced as "kuruvi".

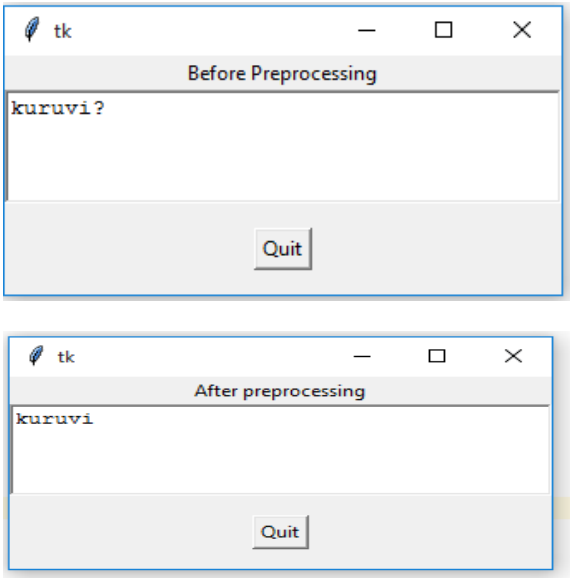

Fig. 5 and Fig. 6. Gui for text pre-processing of the word

C. *Input Text Segmentation And Unicode Conversion*

To produce intelligible speech output, it is required to include all possible sounds of each letter. In such cases, the corpus grows up to 5000 units approximately in syllable based TTS system. Thus the proposed system involves letter level segmentation of units which reduces the size of the corpus to onetenth of the existing size.

The speech corpus size of the letter based synthesis system is approximately equal to the number of letters in Tamil language but it will produce a high naturalness and intelligibility speech. In the proposed system input will be given in English and based on the given input, corresponding Unicode for the Tamil Language is assigned and then letter level segmentation is performed for the Tamil language[3].

Table-II : Unicode letters for Tamil Language

Unicode	Tamil Character
0B85 to 0B94	அ – ஔ
0B95 to 0BA9	க – ன
0BBE to 0BCC	ா- ௌ
0BCD	◌்

The Unicode coded character set is coded in the form of integer values, that has been referred as Unicode scalar values(USVs)[8]. Accordingly, Unicode code points are symbolized in hexadecimal notation of four digits in minimum and preceded with "U+"; so, for example, "U+0345", and "U+20345". Also all the leading zeroes above four digits are suppressed [2]. For the Tamil language, the Unicode ranges from U+0B80 to U+0BFF. It includes Independent vowels ranging from 0B85 to 0B94, Consonants ranging from 0B95 to 0BB9, Dependent vowel signs ranging from 0BBE to 0BCC and others for various signs. Compound characters can be generated by concatenating the Unicode of consonants and Unicode of dependent vowels.

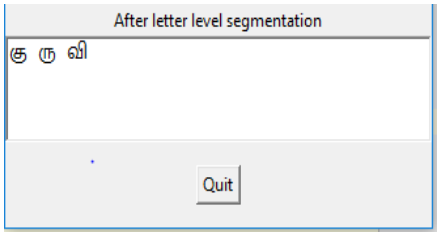

Fig. 7. Gui for letter level segmentation of the word

D. Selection Of Speech Units Unit Selection

Stress of the syllables plays a major role in achieving co-articulation. To achieve Co-articulation, the corpus has to be tuned to store some speech units that help to achieve a fluent speech output [4]. Process involves identifying the stress of letters at the end of first word and at the beginning of second word and selecting units based on the stress. Hence the matched audios for the letters have been fetched using this unit selection method[7].

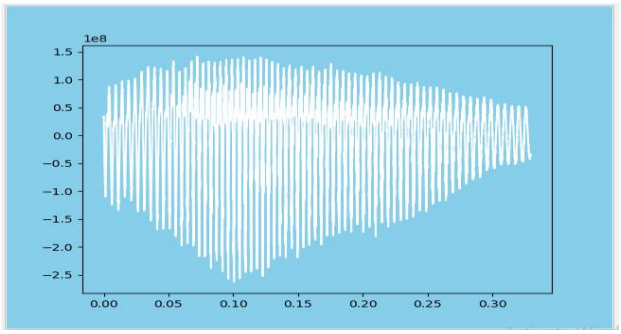

Fig. 8. Waveform Representaion for the letter " ku"

E. Concatenation Of Wav Files Using Fraction Based Concatenation

The concatenation technique for the fraction duration-based waveform specifies the exact fractions of the audio parts [1]. In speech starting point of the vowel in tamil language known as the vowel onset point (VOP) is established. Via manual wave analysis, the consonant sections are known as vowels are produced with high energy when compared to the consonants. Then the speech waveform is concatenated as single speech unit and produced as output.

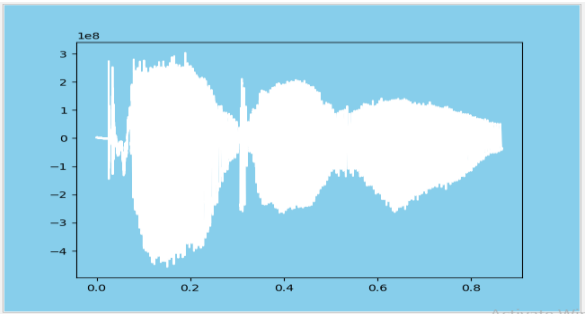

Fig. 9. Waveform Representaion for the letter " kuruvi"

F. Smoothing

To smooth the segment-to-segment transition to achieve consistent production as human speech.

A. Optimal Coupling

The boundaries of speech segments are usually fixed in Concatenative synthesis, but the optimal coupling technique allows the boundaries to shift to give the accurate boundary fit with neighbouring segments. A mismatch measure is checked at every possible segment boundaries till the closest match is found out. Any form of measure can be used, for the improving the spectral quality by using a spectral discontinuity measure is correct. The measures used are Melfrequency Cepstral Coefficients (MFCC) and the auditory-neural based measure (ANBM). If the provided syllable unit is the be-ginning of the word then the boundary will be set from 5n/6th position to n, where n is called as the length of the syllable unit. Or otherwise the boundary will be set to n/3rd position from the beginning.

VI. RESULTS AND DISCUSSION

Performance of speech is defined by the factors on voice quality, naturalness, and intelligibility. Different experiments were conducted with different people under a different set of conditions to check the naturalness and intelligibility of the output obtained using Fraction based waveform concatenation and optimal coupling technique [1]. The detailed experiments and results are briefed below

A. Evaluation

The speech quality test is to assess the quality of synthesized speech output so that the system. The results are analyzed using Mean Opinion Score (MOS) and Precision Recal l measure.

a. Mean Opinion Square(MOS)

The Mean Opinion Score (MOS) provides a number for the distinguished quality after compression of the obtained information. The MOS value is represented as a single number between the range of 1 and 5, where 1 is the lowest distinguished quality of sound and 5 is the highest distinguished quality of sound.. The Mean Of Square value is created by mean value of the results of a series of subjective tests. The number of listeners calculates the audio quality of test sentences heard read loudly by both male and female speakers over the communication medium. Comparing with the natural speech, the listeners provided the rate of speech created after the concatenation of audio units. The findings obtained using the signature of these parameters

Table-III. Mean Opinion Square Values

MOS Measure	QUALITY	DESCRIPTION
5	Excellent	Indistinguishable
4	Good	Distinguishable but not annoying
3	Fair	Slightly displeasing
2	Poor	displeasing
1	Bad	Very displeasing

Table-IV. Mean Opinion Square Results

Sl.No	Input Word	Average MOS Value
1	Kuruvi	4
2	Bharathy	3
3	Kili	4
4	Sangamam	2
5	Geetham	3.5

b. Listeners Preference Test

In this test, the intelligibility and naturalness of the synthetic speech output produced from the proposed system are tested by comparing against the natural speech. A mixed collection of 30 words containing both natural and synthesis speech are chosen, and the words were played random of 5 words each to 6 participants and they will predict whether speech is natural or synthetic. If the listener predicts the natural speech as synthetic or the synthetic speech as natural which means the listener can‟t differentiate the natural and synthetic speech, then it will contribute proposed system has good accuracy in naturalness and intelligibility of speech.

Table-V. Listeners Preference Test

Listeners Preference		Predicted	
		Synthetic speech	Natural speech
Actual	Synthetic	5(a)	25(b)
	Natural	20(c)	10(d)

•Overall Accuracy= (b+c) / (a+b+c+d) * 100

= (45/60)*100 = 75%

• Accuracy(only considered the synthetic input)

=(b/b+a) * 100 = 5/30= 83%

VII. CONCLUSION

The speech synthesis system has been developed to produce TTS for Tamil language. Pre-processing of all the words in the developed speech corpus is done to remove the noise. The output produced continuous speech with higher quality. Intelligibility and Naturalness has to be achieved for the entire speech system by using optimal coupling technique. The naturalness of the speech is enhanced by fraction based concatenation technique.

REFERENCES

1. Soumya Priyadarsini Panda, Ajit Kumar Nayak,“A waveform concatenation technique for text-to-speech synthesis”, International Journal of Speech Technolology, 2017, pp.959–976.
2. M.Karthikadevi and Dr.K.G.Srinivasagan, “The Development of Syllable Based Text to Speech System for Tamil language”, proceedings of International Conference on Recent Trends in Information Technology, vol.14, 2014, pp.167.
3. Dr. Rathinavelu Arumugam, Ms.Jayashree Chelladurai, Ms.Geetha Subburaj, “Tamil Speech Synthesis System: Letter Level Segmentation and Pitch Based Concatenation”, CSI Journal of Computing, Volume: 3 e-ISSN: 2277-7091,2017, pp: 46-55
4. J.Sangeetha, S. Jothilakshmi, S.Sindhuja, V.Ramalingam, “Text to Speech Synthesis system for tamil” , International Journal of Emerging Technology and Advanced Engineering , Vol 3, 2013, Special Issue 1.
5. Deepika Singh, Parminder Singh, “Removal of Spectral Discontinuity in Concatenated Speech Waveform”, International Journal of Computer Application, Volume 53,2012, pp.16
6. S. Saraswathi, R. Vishalakshy, “Design of Multilingual Speech Synthesis System”, Intelligent Information Management, vol 2, 2010, pp. 58-64
7. Sang-Jin Kim and Kyung Ae Jang, “A New Spectral Smoothing Algorithm for Unit Concatenating Speech Synthesis”, IEEE Conference on information and Technology, vol. 3809, 2005, pp. 550-556.
8. “The Unicode Standard 8.0”, Available at: http://www.unicode.org/charts.
9. Hunt,A. and Black, A.”Unit selection in a concate native speech synthesis system using a large speech database”, Proceedings of International conference on acoustics,speech and signal processing 96, Atlanta, Georgia, Volume- 1, pp 373-376
10. N.P.Narenda , K. Sreenivasa Rao,Krishanendu, Ghosh,Ramu Reddy Vempad,”Development of syllable-based text to speech synthesis system in Bengali”, Springer,Int J.Speech Technol. No.14, 2011,pp. 167-181.
11. Youcef TABET , Mohamed BOUGHAZI, “Speech Synthesis Techniques. A Survey”, 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA)IEEE, 2011

AUTHORS PROFILE

Saranya S, PG Scholar, Department of Computer Science and Engineering, Dr.Mahalingam College of Engineering and Technology(MCET), Pollachi. She completed her Diploma in Computer Engineering at P A Polytechnic College, Pollachi. She completed her Bachelor of Engineering in Thaigarajar College of Engineering , Madurai. she has 6 months of Industrial Experience. She is doing her project in Human Computer Interaction.

Dr.A.Rathinavelu, Principal, Dr,Mahalingam College of Engineering and Technology(MCET), obtained his Ph.D from NIT, Trichy. He completed his Master of Technology in Edith Cowan University, Perth, Western Australia and Bachelor of Engineering at IRTT ,Erode. He has 6 years of Industrial experience in Engineering college. He is working in MCET since 1999 in various capacities. He is a Professor in the department of Computer Science and Engineering. His research area includes Human Computer Interaction, Web Engineering, Speech Processing and Mobile Application Development.

Jayashree C, Assistant Professor, Department of Computer Science and Engineering, Dr. Mahalingam College of Engineering and Technology. She completed her Master of Engineering in Computer Science in 2016 and Bachelor of Engineering form Anna University Chennai in 2012. She has Industry experience for one year and teaching experience of more than three years. Her current areas of research are Speech Processing and Data Analytics.