ISSN: 2321-7782 (Online)
Volume 3, Issue 5, May 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Malayalam Text To Speech Synthesis with Emotions

Aiswarya T S[1]
Signal Processing
College of Engineering Cherthala
Alappuzha, Kerala, India

Jayadas C K[2]
Signal Processing
College of Engineering Cherthala
Alappuzha, Kerala, India

Abstract: Many text-to-speech synthesizers for Indian languages have used synthesis tech- niques that require prosodic models for good quality synthetic speech. However, due to unavailability of adequately large and properly annotated databases for Indian languages, prosodic models for these synthesizers have still not been developed properly. With inadequate prosodic models in place, the quality of synthetic speech generated by these synthesizers is poor.The recent development in text to speech has been switched to concatenative synthesis, either using original speech segments or parametric synthesis. The former TTS system gives a better quality output since they use the original speech segment for concatenation. There are a number of different other techniques for speech generation like PSOLA, TDPSOLA, EMBROLA etc.The TTS uses diphone like segments (partneme) as the basic units for concatenation.

Keywords: prosody, concatenative synthesis, PSOLA, EMROLA.TTS

I. Introduction

One of the most distinct features of human civilizations that have evolved along with them over time is probably the language. Speech holds an important role in this evolution by not only being the spoken form of a language but also the most efficient way of communication. Probably because of its profound influence on day to day human lives, speech has been a subject of constant research for centuries. Various aspects of speech - recognition, synthesis, understanding and identification have been studied to see how these tasks, that humans do effortlessly, can be done by machines. The goal of text to- speech synthesis (TTS) is the automatic conversion of unrestricted natural language sentences in text form to a spoken form that closely resembles the spoken form of the same text by a native speaker of the language. This field of speech research has witnessed significant advances over the past decade with many systems being able to generate a close to natural sounding synthetic speech. Research in the area of speech synthesis has been fueled by the growing importance of many new applications. These include information retrieval services over telephone such as banking services, public announcements at places like train stations and reading out manuscripts for collation. Speech synthesis has also found applications in tools for reading emails, faxes and web pages over telephone and voice output in automatic translation systems. Special equipment for the physically challenged, such as word processors with reading-out capability and book-reading aids for visually challenged and speaking aids for the vocally challenged also use speech synthesis [1].

In most of the TTS systems the improvements of quality of speech synthesis are aimed at stimulating neutral speech, reading a neutral text in a neutral speaking style. Hence the synthetic voice is rather monotonous. The inclusion emotional effects can results in expressive speech, decreasing the monotony of synthetic speech .The emotional aspects are manifested in speech mainly as change in prosodic parameters: pitch, duration and intensity. Prosody is the intonation, rhythm, and lexical stress in speech. The acoustical changes caused due to emotions are language dependent, so it is necessary to perform analysis for each language. A lot of researches are going on in the field of emotional analysis and recognition. In Malayalam, emotional speech analysis has not been carried out previously. Mainly input text from an image to a speech synthesis system consists of a character recognizer and TTS system .Resynthesis of emotional speech from neutral speech requires another system.

© 2015, IJARCSMS All Rights Reserved

A. Nature of Language

Malayalam, like most of the other Indian Languages, is a phonetic language having a written form that has direct correspondence to the spoken form. Each character corresponds to a syllable, which has an invariant pronunciation irrespective of the context in which it occurs (with only one or two exceptions). Due to the one to one correspondence between letters and phonemes, framing rules for extracting phonemes from words is comparatively uncomplicated. The phonemes (called varnams in Sanskrit) are divided into two types: vowel phonemes (swara varnam) and consonant phonemes (vyanjan varnam). They together broadly constitute the Varnamala (alphabet set). The orthographic representation of these varnams is done in a systematic way. While Swara Varnam is self-powered and it is not dependent on any other element, the Vyanjana Varnam however, needs an addition of Swara Varnam to compose a syllabic entity. The Swara Varnam also called nadi (one that produces sound) are all voiced sounds while the Vyanjan Varnam also called as swasi (one that producing swasam or flow of air) can be voiced or unvoiced. The combination of consonant phoneme and a vowel phoneme produces a syllable (akshara). The phonemes when combined as C...C + V or only V form complete phonetic cluster. In other words each letter is formed of either a vowel or a vowel in combination with one or more consonants. The Varnamala or alphabet set is phonetically structured. The vowels and consonants are separately grouped and systematically arranged. The set of 16 vowels forms the first row of varnamala followed by stop consonants. The phonemes are categorized according to the method of speech production and articulation. The row wise arrangement is according to the manner of articulation, whereas the column wise arrangement is according to the method of speech production. The fricatives, semivowels etc. are grouped separately as a miscellaneous set. The phonetic nature of the language and the systematic categorization of the alphabet set can be effectively used for analysis and modeling.

II. Related Work

Optical character recognition (OCR)[2][7] is a process of automatic computer recognition of characters in optically scanned and digitized pages of text. OCR is one of the most fascinating and challenging areas of pattern recognition with various practical application potentials. Depending on versatility, robustness and efficiency, the commercial OCR systems can be divided into four generations. The first generation systems can be characterized by the constrained letter shapes which the OCR read. Such machines appeared in the beginning of the 1960.The recognition method was template matching. The next generation is characterized by the recognition capabilities of a set of regular machine printed characters as well as hand-printed characters. At the early stages, the scope was restricted to numerals only. Such machines appeared in the middle of 1960 to early 1970.The methods were based on the structural analysis approach. The third generation can be characterized by the OCR of poor print quality characters, and hand -printed characters for a large category character set. Commercial OCR systems with such capabilities appeared roughly during the decade 1975 to 1985.The fourth generation can be characterized by the OCR of complex documents intermixing with text, graphics, table and mathematical symbols, unconstrained hand written characters, color document, low-quality noisy documents like photocopy and fax, etc. Traditionally, pattern recognition techniques are classified as template and feature-based approach. In the template-based approach, an unknown pattern is superposed directly on the ideal template pattern and the degree of correlation between the two is used for the decision about classification. Early OCR systems employed template based approach, but modern system combines with feature based approaches to obtain better results.

Synthesized speech can be produced by several different methods. All of these have some benefits and deficiencies that are discussed in this and previous chapters. The methods are usually classified into

Three groups:

» Articulatory synthesis, which attempts to model the human speech production system directly

- Formant synthesis, which models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model.
- Concatenative synthesis, which uses different length prerecorded samples derived from natural speech.

The formant and concatenative methods are the most commonly used in present synthesis systems. The formant synthesis was dominant for long time, but today the concatenative method is becoming more and more popular. The articulatory method is still too complicated for high quality implementations, but may arise as a potential method in the future.

III. PROBLEM FORMULATION

In Malayalam, emotional speech analysis has not been carried out previously. In recent years OCR system has received considerable attention because of the tremendous need for digitization of printed document. The goal of OCR is to classify optical patterns corresponding to alphanumeric or other characters. An OCR system for printed text documents in Malayalam, segments the scanned document images into text line words and further characters. The scanned image of a printed Malayalam text is the input to the system and the output is the editable computer file containing the text data in the printed page. Segmentation and feature extraction are the most important phases involved in the system. There are many OCR systems available for handling English documents, however there are many not reported effort for Indian languages. The problem area in speech synthesis is very wide. There are several problems in text preprocessing, such as numerals, abbreviations, and acronyms. We need to know that the sentence ends after a full stop (.) and not between abbreviations. It is somewhat easy to tokenize a word with help of full stop as most of the sentences will be ending with full stop. But there are some other cases where it ends with semicolon or some other punctuation like previous case. This problem can be solved by expanding the abbreviation and removing the unwanted punctuation. All the Malayalam abbreviations cannot be expanded, because some mostly used abbreviations are stored in a separate database. When certain abbreviation comes in the text, then it will search in the database for that abbreviation. If that abbreviation is present the system will replace the text, if not it will be leaving the original text as it is. It is difficult to add all the abbreviations in the database, so most commonly used abbreviations are used. The unwanted punctuation like (: , ; ‘) etc. are to be removed from the given paragraph to avoid confusion and not to give any disturbance in the naturalness of the speech The second step in text normalization is normalizing non-standard words. Non standard words are tokens like numbers or abbreviations, which need to be expanded into sequences of Malayalam words before they can be pronounced. Ambiguity is the main problem with the non-standard words. For example, the number 1900 can be spoken in at least three different ways, depending on the context

ആയിരത്തി തൊള്ളായിരം

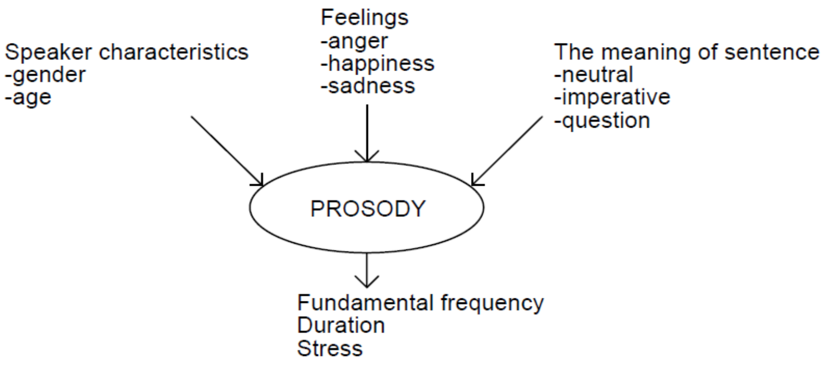
പത്തൊൻപ്ത് നൂറ്

ഒന്ന് ഒൻപത് പൂജ്യം പൂജ്യം

And the same letter which has two sounds is identified by the conditions. For eg:

നനഞ്ഞു=ന+ന+ഞ്ഞു

Here first na() and second na() has different pronunciation. First letter is the dental nasal and second letter is the aiveolar nasal. Finding correct intonation, stress, and duration from written text is probably the most challenging problem for years to come. These features together are called prosodic or supra segmental features and may be considered as the melody, rhythm, and emphasis of the speech at the perceptual level. The intonation means how the pitch pattern or fundamental frequency changes during speech. The prosody of continuous speech depends on many separate aspects, such as the meaning of the sentence and the speaker characteristics and emotions. Unfortunately, written text usually contains very little information of these features and some of them change dynamically during speech. However, with some specific control characters this information may be given to a speech synthesizer. It is observed that the development of TTS in Indian languages is a difficult task, especially for Malayalam, in which same letters is pronounced in multiple ways. Success of Malayalam TTS depends not only on addressing of above said issue but also in corporating of regional variation in speaking of Malayalam.



IV. Proposed Method

Proposed OCR

In recent years OCR system has received considerable attention because of the tremendous need for digitization of printed document. The goal of OCR is to classify optical patterns corresponding to alphanumeric or other characters. An OCR system for printed text documents in Malayalam, segments the scanned document images into text line words and further characters. The scanned image of a printed Malayalam text is the input to the system and the output is the editable computer file containing the text data in the printed page. Segmentation and feature extraction are the most important phases involved in the system.

The proposed system involves four phases, scanning of the image, pre-processing, feature extraction and recognition

a) Scanning the image

The document is scanned by any standard scanners. Lower resolution results in poor performance of the system and misidentifications. A scanned image will be any one of format jpeg, bmp or tiff.

b) Preprocessing

The scanned image contains noise, and these should be preprocessed to remove them from the image. Secondly the characters should be individually extracted from the original image. The preprocessing stage in character recognition consists of Removal of noise if any, Binarization of the Image, Separation of words, Identification of line and space in the passage and resizing to a standard size.

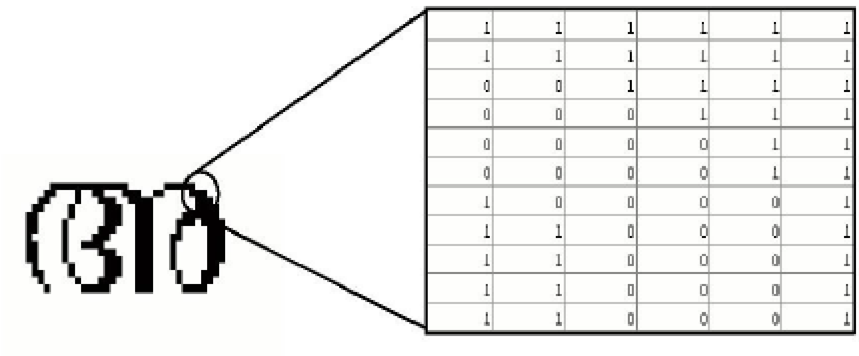
c) Removal of Noise

Noises in an image are of two types. They are Gaussian Noise and Salt and Pepper Noise. Image filtering removes the Gaussian noise. Even though salt and pepper noise can be removed, their effect will be there throughout the process. Filtering done here is by spatial domain (by using filter masks). Frequency domain filtering leads to the loss of data when reconstructing.

In the filtering stage, a spatial domain filter is used for the removal of noise. The isolated and dilated points in the image are removed, where a clear image is obtained after filtering.

d) Binarisation

In a binary image, only two levels will be there which are 0 and 1. They are also called as logical images Here the value 0 corresponds to black pixel and 1 corresponds to a white pixel. A binary image with its eight connected pixels of an image.



1	1	1	1	1	1
1	1	1	1	1	1
0	0	1	1	1	1
0	0	0	1	1	1
0	0	0	0	1	1
0	0	0	0	1	1
1	0	0	0	0	1
1	1	0	0	0	1
1	1	0	0	0	1
1	1	0	0	0	1
1	1	0	0	0	1

Fig. Binarisation

In the pre-processing stage the scanned image obtained is converted to binary image for easy analysis and to get the information about the lines and spaces between the characters. To convert a grayscale image to binary, a threshold level of the grayscale image is found out. This threshold level corresponds to the gray level threshold of the grayscale image. A hard limiter is set with this gray level threshold.

e) Separation of Characters

Characters can be separated by vertical and horizontal scanning of the acquired image. The horizontal scan is done for the separation of lines from a document. The scanning starts from the top by vertical path and finds the gap present and then separates the pixels above the gap and vertical scanning is done.

The vertical scanning involves the scanning of image from horizontal scanning process and then determines the gap between the words and separates it. Thus the words can be separated from a document.

f) Combinational letters in Malayalam

Horizontal and vertical scanning only separates the characters by identifying the gap between them as explained earlier. But in a Malayalam character set there are different combinational letters that cannot be separated by this process. One such combination is pronounced as ’ki’, where there is no gap between the two letters. A separate algorithm for of this should be developed to separate these types. One of the best methods labeling the image. Labeling is done by the four and eight connected pixels of an image. The connected component is done only for a binary image with pixel varies from ’0’ and ’1’. The labeling process will be done for the connected components for the value ’1’. So the image is converted to complement image, where the back ground will be black and foreground will be white. A four connected components in an image f(x, y) is four pixels in the adjacent boundaries of a pixel (i,j).i.e.(i-1,j), (i,j+1), (i+1, j), (i,j-1).Similarly for the pixel 8-connected component is defined as (i-l,j),(i-1,j+l),(i,j+l),(i+1,j+l),(i+1,j),(i+1,j1),(i,j1),(i-1,j-1).Here the first letter is labeled as ’1’ and the second letter is labeled as 2.

g) Resizing

The character should be in a standard form for the purpose of recognition. Therefore it should be resized to a standard size. Bilinear Interpolation technique is used here. It maps each pixel in the new image back to a point in the old image where the four nearest pixels to this point are used to assign the value to the new pixel. These four pixels are used in a calculation that is linear in x and linear in y but not linear in both together, in other words, it is bilinear. The characters are resized to a standard size of 32 x 64, by the above-mentioned resizing algorithm. Resizing the image will obtain font invariance during the recognition process.

h) Feature Extraction

For the extraction of the features from the character the Daubechies (db4) wavelet is used. The wavelet transform of a signal s is the family C (a, b), which the depends on two indices a and b. The wavelet decomposition consists of calculating a ”resemblance index” between the signal and the wavelet located at at the position b and of scale a. If the index is large, the further resemblance is strong, otherwise it is slight. The indexes C (a, b) are called coefficients.

i) Recognition

Recognition is accomplished by using neural networks. The most commonly used family of neural networks for pattern classification tasks is the feed- forward network, which includes multi layer perceptron and Radial-Basis Function (RBF) networks. These ta networks are organized into layers and have unidirectional connections between the layers. Another popular network is the Self- Organizing Map (SOM), or Input from the Kohonen-Network, which is mainly used for data Extracted Feature clustering and feature mapping. The learning process involves updating network architecture and connection weights so that a network can efficiently perform a Calculate specific classification/clustering task. The increasing popularity of neural network models to solve pattern recognition problems has been primarily due to their low dependence on domain-specific knowledge (relative to model-based and rule-based Weight approaches) and due to the availability of efficient learning algorithms for practitioners to use.

Having localized and coded acquired image that Calculate corresponds to the character, the final task is to decide if this character code matches a previously stored character code. Here we are using a feed forward back propagation neural network for that purpose. In general, competitive learning neural networks are used for fast learning mechanism. But their performance is easily affected by initial weight vectors. A feed forward back propagation network, overcomes this difficulty by automatically initializing the weight vectors.

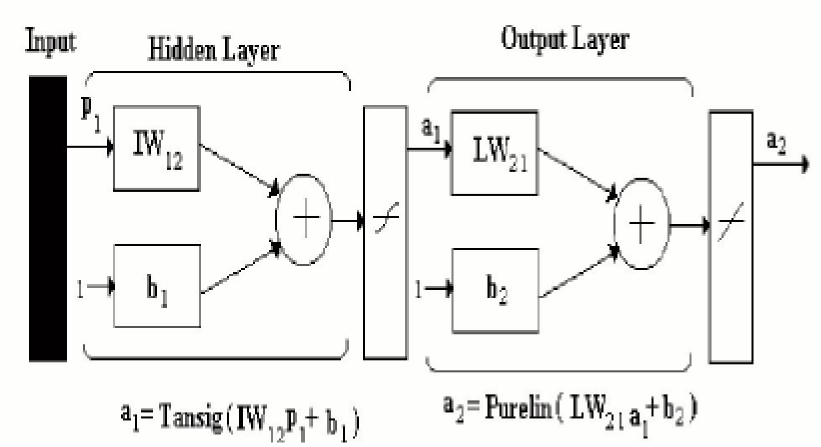


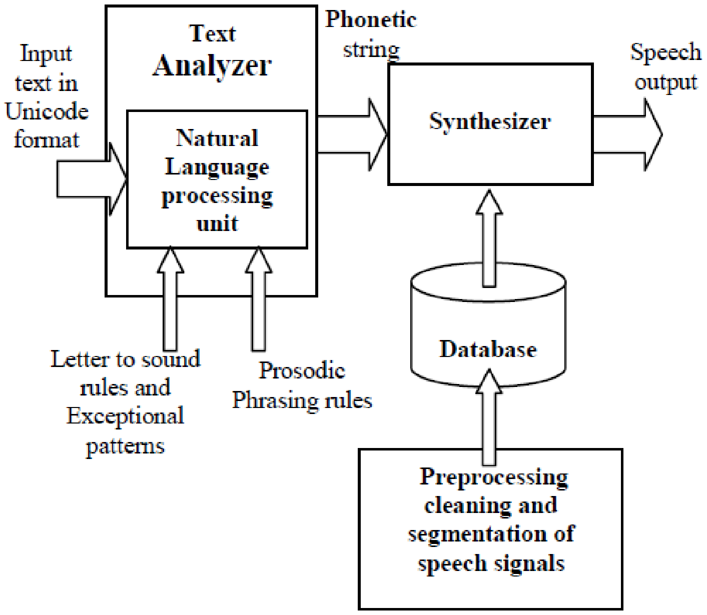
Fig.Feed forward Back Propagation Network

Proposed TTS System

The main modules of a TTS system are the text analyzer module and the synthesizer. The text analyzer along with the Natural Language Processing (NLP) unit generates the phonetic string, in the format that can be processed by the synthesizer. The NLP module determines the accuracy/correctness of the units for concatenation. The synthesizer reads this input; identify the units (partneme –part of phones) to be concatenated. The synthesizer then selects the partneme from database and concatenates the selected partneme to generate the speech output. While concatenation the synthesizer applies some signal processing to adjust the pitch, and duration. The synthesizer also takes care to reduce the distortion at the concatenation point by ensuring Epoch synchronous concatenation.

© 2015, IJARCSMS All Rights Reserved

The text analyzer accepts the input string, normalizes the string with language specific modifications, apply language specific rules and generate the phonetic string. Synthesizer is the core part which converts the string to wave with fairly good naturalness.


Input text in Unicode format
Text Analyzer
Natural Language processing unit
Phonetic string
Synthesizer
Speech output
Database
Letter to sound rules and Exceptional patterns
Prosodic Phrasing rules
Preprocessing cleaning and segmentation of speech signals

A. The Text Analyzer Module

The text analyzer reads the input text and cleans the input text. Cleaning is done to remove all unwanted characters, common typing errors etc. Cleaned text is then parsed to the NLP module to extract abbreviations, numbers etc. and to apply LTS rules. Abbreviations are replaced with its expansion. In the current implementation frequently used abbreviations are handled. The numbers in figures must be converted to words before applying to synthesizer. The normalization of numbers in Malayalam is not simple as other Indian languages like Hindi, Bangla etc. The numbers usually appear in figure along with the suffix patterns. The text analyzer identifies the number and suffix, replace the number and suffix with its equivalent expansion after applying agglutination rules.

B. Speech Database

The speech database is created using partnemes which are diaphone like units, covering the all possible combinations for the language. The advantage of using partnemes as the basic unit is the simplicity of introducing intonation and prosodic rules into the synthesized speech signals. The creation of partneme database involves recording of the nonsense words to get all possible combinations in neutral mode, with almost constant pitch and amplitude.

C. Synthesizer

The synthesizer identifies the segment to be concatenated from the phonetic string which represents the actual pronunciation. The synthesizer has a token generation module, which generates the token using token generation rules as below, which is used to identify the partneme for concatenation. The speech generation module from the header extract the starting address, reads the data from the database, concatenate the segments by ensuring minimal distortion at the concatenation point. Joining of speech segments is done at epoch synchronous points to ensure minimal distortion at the concatenation points The sample for token generation rule is given below. Tokens are generated based on the succeeding and preceding phones. These rules are language specific. These tokens correspond to indexing of segmented partneme voice signals in the speech database header.

Resynthesis of Emotional Speech from Neutral Speech

PSOLA For modification of the input speech, only prosody modifications were performed using the TD-PSOLA algorithm as implemented in the Praat software[20]. The modifications were performed on the voiced (V) and unvoiced regions (U) of utterances by scaling the original utterance values by the factors listed below. The voiced and unvoiced region boundaries were automatically detected using the Praat software The PSOLA (Pitch Synchronous Overlap Add) method was originally developed at France Telecom (CNET). It is actually not a synthesis method itself but allows prerecorded speech samples smoothly concatenated and provides good controlling for pitch and duration, so it is used in some commercial synthesis systems, such as ProVerbe and Hadifix (Donovan 1996).

There are several versions 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). The basic algorithm consist of three steps (Charpentier et al. 1989, Valbret et. al 1991). The analysis step where the original speech signal is first divided into separate but often overlapping short-term analysis signals (ST), the modification of each analysis signal to synthesis signal, and the synthesis step where these segments are recombined by means of overlap-adding. Short term signals xm(n) are obtained from digital speech waveform x(n) by multiplying the signal by a sequence of pitch-synchronous analysis window hm(n): xm(n) = hm(tm -n)x(n) where m is an index for the short-time signal. The windows, which are usually Hanning type, are centered around the successive instants tm, called pitch-marks. These marks are set at a pitchsynchronous rate on the voiced parts of the signal and at a constant rate on the unvoiced parts. The used window length is proportional to local pitch period and the window factor is usually from 2 to 4. The pitch markers are determined either by manually inspection of speech signal or automatically by some pitch estimation methods (Kortekaas et al. 1997). The segment recombination in synthesis step is performed after defining a new pitch-mark sequence.

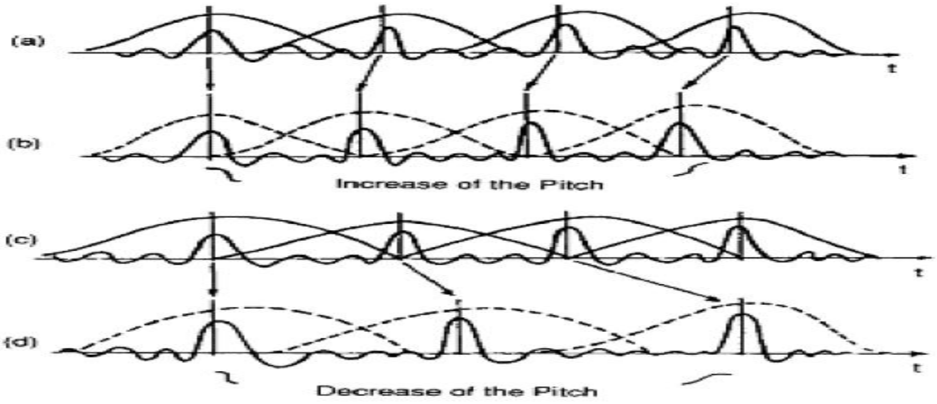


Fig.Pitch modification of a voiced speech segment

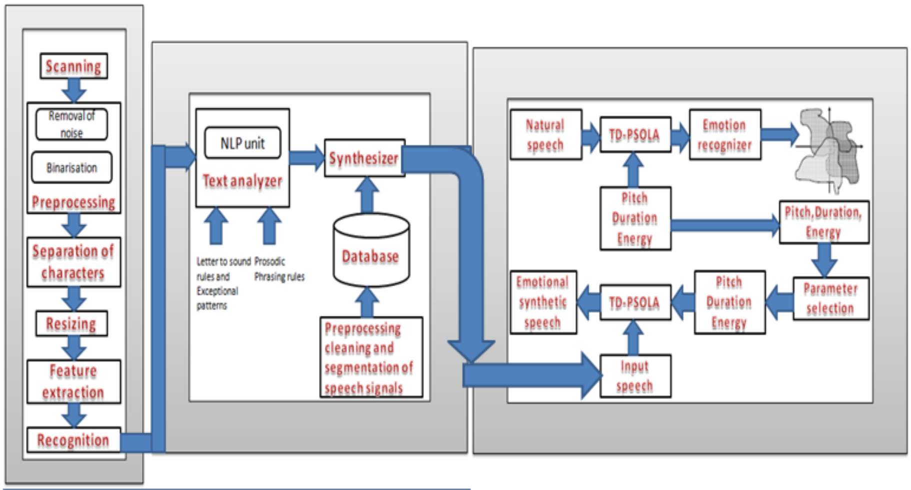


Fig. Proposed Malayalam TTS With Emotions

V. Conclusion

Text to Speech synthesizer of Malayalam can be implemented by using OCR, TTS and psola method. There is prior research done regarding this topic. When increasing the features selected the OCR perfect will more accurate.

References

1. Daniel Jurafsky and James H.Martin, Speech and Language Processing
2. Mohamed Cheriet,Nawwaf Kharma,Character Recognition Systems,Persons
3. Zhengqiang Guan;Bo Yang;Pengfei Li; " Study on Template-Based Coding Method of Binary Ocr" IEEE,vol.2
4. Lucas, S.M., "High performance OCR with syntactic neural networks," Artificial Neural Networks, 1995., Fourth International Conference on , vol., no.,pp.133-138, 26-28 Jun
5. Simon Haykin,Neural Network and Learning Machines,Person
6. Malyan, R.R.; Sunthankar, S.; Teranchi, H.; Yeghiazarian, A., "Perception of multiauthor handprinted text," Character Recognition and Applications, IEE Colloquium on , vol., no., pp.8/1-8/3, 2 Oct 1989
7. Lynch, M.R.; Rayner, P.J., "Optical character recognition using a new connectionist model," Image Processing and its Applications, 1989., Third International Conference on IEEE, vol., no., pp.63-67, 18-20 Jul 1989
8. Mori, S.; Suen, C.Y.; Yamamoto, K., "Historical review of OCR research and development," Proceedings of the IEEE , vol.80, no.7, pp.1029-1058, Jul 1992
9. MAajitha Ali Abed Hamid ali Abed Simplifying Handwritten Characters Recognition Using a Particle Swarm Optimization Approach EUROPEAN ACADEMIC RESEARCH, VOL. I, ISSUE 5/ AUGUST 2013 ISSN 2286-4822
10. Elliman, D.G., "Peeling potatoes with a cheese grater [handwritten document OCR] ," Handwriting Analysis and Recognition: A European Perspective, IEE European Workshop on , vol., no., pp.14/1-14/4, 12-13 Jul 1994
11. Tanprasert, C.; Koanantakool, T., "Thai OCR: a neural network application," TENCON '96. Proceedings. 1996 IEEE TENCON. Digital Signal Processing Applications , vol.1, no., pp.90-95 vol.1, 26-29 Nov 1996 Department of Electronics Engineering 40 College of Engineering, Cherthala CHAPTER 5. FUTURE WORKS
12. Nagy, G.; Prateek Sarkar, "Document style census for OCR," Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on , vol., no., pp. 134-147, 2004
13. Shaolei Feng; Manmatha, R., "A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books," Digital Libraries, 2006. JCDL '06. 73 Proceedings of the 6th ACM/IEEE-CS Joint Conference on , vol., no., pp.109-118, June 2006
14. Ainsworth, W., "A system for converting English text into speech," Audio and Electroacoustics, IEEE Transactions on , vol.21, no.3, pp. 288-290, Jun 1973
15. Fushikida, Katsunobu; Mitome, Yukio; Inoue, Yuji, "A Text to Speech Synthesizer for the Personal Computer," Consumer Electronics, IEEE Transactions on , vol.CE- 28, no.3, pp.250-256, Aug. 1982
16. Leija, L.; Santiago, S.; Alvarado, C., "A system of text reading and translation to voice for blind persons ," Engineering in Medicine and Biology Society, 1996. Bridging Disciplines for Biomedicine. Proceedings of the 18th Annual International Conference of the IEEE , vol.1, no., pp.405-406 vol.1, 31 Oct-3 Nov 1996
17. Dey, S.; Kedia, M.; Basu, A., "Architectural Optimizations for Text to Speech Synthesis in Embedded Systems," Design Automation Conference, 2007. ASPDAC '07. Asia and South Pacific , vol., no., pp.298-303, 23-26 Jan. 2007
18. MAbdul Rahiman,M S Rajasree,"Printed Malayalam Character Recognition Using Back-propagation Neural Networks" 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009
19. Arun Gopi, Shobana Devi P, Sajini T, Bhadran V K,"Implementation of Malayalam Text to Speech Using Concatenative Based TTS for Android Platform" 2013 IEEE
20. Murtaza Bulut, Sungbok Lee and Shrikanth Narayanan,"Recognition for Synthesis: Automatic Parameter Selection for Resynthesis of Emotional Speech from Neutral Speech" 2008 IEEE