Abstract - Several prototypes and fully operational systems have been built based on different synthesis techniques. This paper describes the Text to Speech Synthesizer with prosody feature for generating natural human conversation. It also deals with forward parsing technique for sentiment analysis of the TTS system in an efficient way. The prosody feature plays an important roll to generate the TTS output just like human voice. To produce the output of TTS in the same form as if it is actually spoken the prosody feature allows the synthesizer to vary the pitch of voice. The forward parsing is utilized to compute the sentiment in the text and the sentiment based output will be generated.

Keywords - Text to speech synthesizer (TTS), Forward parsing, prosody feature

I. INTRODUCTION

Over the past years there has been a great development in speech technology. Among the applications of speech technology, the automatic speech production, which is referred to as text-to-speech (TTS) system is the most natural sounding technology. Text-to-speech synthesis is the process of converting ordinary orthographic text into acoustic signal which is indistinguishable from human speech[1]. Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware products.

A (TTS) system converts normal language text into speech, other systems render symbolic linguistic representations like phonetic transcriptions into speech[2]. To implement complex waves of physical, phonetic effects that is being applied to express speculation, attitude and attention as a parallel channel in our daily speech communication is carried out by prosody feature. From listener aspect prosody contains systematic perception and recovery of speaker intention based on relative volume, rate of vocal fold cycle as function of time (pitch), phoneme duration and time.

Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or di-phones provides the largest output range, but may lack clarity.

For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output

II. GENERAL ARCHITECTURE OF TTS SYSTEM

A text-to-speech system (or "engine") is composed of two parts: a front-end and a back-end. The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This process is often called text normalization, pre-processing, or tokenization[3].

The front-end then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front-end. The back-end often referred to as the synthesizer—then converts the symbolic linguistic representation into sound. In certain systems, this part includes the computation of the target prosody

![Figure 1: TTS System Architecture](image)

(pitch contour, phoneme durations), which is then imposed on the output speech.

a) Pre-processing

Possible identification of text genre, character encoding issues, possible multi-lingual issues.
The feature of unit selection synthesis to preserve the features of the recorded speech very well has been exploited by [9] for the synthesis of emotional speech. For each of three emotions (anger, joy, and sadness), an entire unit selection database was recorded by the same speaker. In order to synthesis a given emotion, only units from the corresponding database are selected. The emotions in the resulting synthesized speech are well recognized (50-80%).

IV. PROSODY GENERATION

The various elements of prosody generator in TTS system has been visualized in Figure 1. The parsed text with a phoneme string is the input of prosody generator. The duration of each phoneme and the pitch contour is delivered by the prosody generator. Before applied input to theProsody generator the input is parsed and converted in to phonemes based on the key stokes involved in the characters present in the input. The duration and pitch of each phoneme depends on the content and context of the text. In the context, the mood of conversation is happy, then pitch of word is changed accordingly to allow listener to understand happy mood of content. So prosody has an important roll in guiding listener recovery of the basic messages.
Prosody: Prosody depends not only on the linguistic content of a sentence. Different people generate different prosody for the same sentence. Even the same person generates a different prosody depending on his or her mood. The speaking style of the voice in Figure 1, can impart an overall tone to a communication. Examples of such global settings include a low register, voice quality (falsetto, creaky, breathy, etc), narrowed pitch range indicating boredom, depression, or controlled anger, as well as more local effects, such as notable excursion of pitch, higher or lower than surrounding syllables, for a syllable in a word chosen for special emphasis. The various parameter which influence the speaking Style are presented in[11,12].

a. Character:

It is an important determining element in prosody, refers primarily to long-term, stable, extra-linguistic properties of a speaker, such as membership in a group and individual personality.. In determines the features such as gender, age, speech defects, etc. affect speech, and physical status may also be a background determiner of prosodic character. Finally, character may sometimes include temporary conditions such as fatigue, inebriation, talking with mouth full, etc. Since many of these elements have implications for both the prosodic and voice quality of include temporary conditions such speech output, they can be very challenging to model jointly in a TTS system.[5,7]

b. Emotion:

The prosody should consider the temporary emotional conditions such as amusement, anger, contempt, sympathy, suspicion, etc are suggested in[11][12]. A large number of high-level factors go into determining emotional effects in speech. Among these are point of view (can the listener interpret what the speaker is really spontaneous vs. symbolic (e.g., acted emotion vs. real feeling); culture-specific vs. universal; basic emotions and compositional emotions that combine basic feelings and effects; and strength or intensity of emotion.

1. Anger, though well studied in the literature, may be too broad a category for coherent analysis. One could imagine a threatening kind of anger with a tightly controlled F0, low in the range and near monotone; while a more overtly expressive type of tantrum could be correlated with a wide, raised pitch range.

2. Joy is generally related with increase in pitch and pitch range, with increase in speech rate. Untrained listener can easy to identify smiling because it raises F0 and formant frequencies.

3. Sadness generally has normal or lower than normal pitch realized in a narrow range, with a slow rate and tempo. It may also be characterized by slurred pronunciation and irregular rhythm.

4. Fear is characterized by high pitch in a wide range, variable rate, precise pronunciation, and irregular voicing.

VI. DURATION ASSIGNMENT

The phoneme durations are characterized by various factors. The general factors are

  a. Semantic and Pragmatic Conditions
  b. Speech rate corresponding to speaker intent, mood and emotion
  c. The use of duration or rhythm to possibly signal document structure above the level of phrase or sentence [13]
  d. The lack of a consistent and coherent practical definition of the phone such that boundaries can be clearly located for measurement Rule based method is one of the most commonly used methods for Duration Assignment. For every phone type this method uses table lookup for minimum and inherent duration. The duration is rate dependent, so all phones can be globally scaled in their minimum duration for faster or slower rates.
The inherent duration is raw material and using the specified rules, it may be stretched or contracted by pre-specified percentage attached to each rule type as specified and then it is finally added back to the minimum duration to yield a millisecond time for a given phone. The duration of phone is expressed as

\[ d = d_{\text{min}} + r(d - d_{\text{min}}) \]

Where \( d_{\text{min}} \) is the minimum duration of the phoneme, \( d \) is average duration of the phoneme and correction \( r \) is given by:

\[ r = \prod_{i=1}^{N} r_i \]

For the case of \( N \) rules applied, where each rule has correction \( r_i \). At the very end, a rule may apply that lengthens vowels when they are preceded by voiceless plosives. The list of rules used for calculating duration as follows:

1. Lengthening of final vowel and following consonant in pre-pausal syllables
2. Shortening of all syllabic segments in non-pre-pausal positions
3. Shortening of syllabic segments if not in a word final syllable
4. Consonant in non-word initial positions are shortened
5. Un-stressed and secondary stressed phones are shortened
6. Emphasized vowels are lengthened
7. Vowels may be shortened or lengthened according to phonetic features of their context
8. Consonants may be shortened in cluster

**VII. PITCH GENERATION**

Since generating pitch contours is an incredibly tedious problem, pitch generation is often divided into two levels, with the first level computing the so-called symbolic prosody and the second level generating pitch contours from this symbolic prosody. This division is somewhat arbitrary since, as we shall see below, a number of important prosodic phenomena do not fall cleanly on one side or the other but seem to involve aspects of both. Often it is useful to add several other attributes of the pitch contour prior to its generation, which is discussed in coming section

**VII.a. Pitch Range:**

Pitch range refers to the high and low limits within which all the accent and boundary tones must be realized: a floor and ceiling, so to speak, which are typically specified in Hz. This may be considered in terms of stable, speaker-specific limits as well as in terms of an utterance or passage.

**VIII. FORWARD PARSING METHOD**

Parsing is an important method for scanning the text. Through parsing we can determine various points such as content of text, context of text, frequency of particular word in the text etc. It is necessary to determine context of text to find out the emotions present in it. The context of the text is not only used to determine the current emotions present in the text and also used to determine the variation in the emotion. Most of the commercial available TTS are based on regular parsing in which the emotion present in the text is generated at the same time when the text is converted and represented in the voice form to the user. This approach followed in current text to speech synthesizers. [14] **Forward Parsing Implementing**

The block diagram for implementing forward parsing is as shown in the figure.

![Figure 3: Architecture for Forward Parsing](image-url)

As shown in the figure 3, a Database is maintained, which contains the keywords and category of emotion to which it belongs. Following types of emotions are handled using the architecture.

Anger, Joy, Surprise, Disgust, Contempt, Pride, Depression, Funny, Sorry, Boredom, Suffering, Shame
The text is scanned and keywords present in the text are compared with the contents of database. The comparison will determine the value of emotion. Once the type of emotion is fixed the information is supplied to the composer, which then composes the wave file based on value of emotion[16]. The value of emotion is changed based on intensity of emotion in the text. For example if the text is **Magizchiaga Irukiren**: Then the intensity of emotion is normal and will be represented by <+>. **Figure 4: Magizchiaga Irukiren**

**Migavum Magizchiaga Irukiren**: Then the intensity is increase and will be represented by <++>

**Figure 5: Migavum Magizchiaga Irukiren**

The text is scanned and keywords present in the text are compared with the contents of database. The comparison will determine the value of emotion. Once the type of emotion is fixed the information is supplied to the composer, which then composes the wave file based on value of emotion[16]. The value of emotion is changed based on intensity of emotion in the text. For example if the text is **Magizchiaga Irukiren**: Then the intensity of emotion is normal and will be represented by <+>.

**IX. RESULTS AND DISCUSSION**

<table>
<thead>
<tr>
<th>GTTS</th>
<th>STTS</th>
<th>Emotions</th>
<th>Comparative Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1 (5 Members)</td>
<td>Set 1 (5 Members)</td>
<td>Angry</td>
<td>GTTS &gt; STTS</td>
</tr>
<tr>
<td>Set 2 (5 Members)</td>
<td>Set 2 (5 Members)</td>
<td>Sadness</td>
<td>GTTS &gt; STTS</td>
</tr>
<tr>
<td>Set 3 (5 Members)</td>
<td>Set 3 (5 Members)</td>
<td>Joy</td>
<td>GTTS &gt; STTS</td>
</tr>
<tr>
<td>Set 4 (5 Members)</td>
<td>Set 4 (5 Members)</td>
<td>Fear</td>
<td>GTTS &gt; STTS</td>
</tr>
<tr>
<td>Set 5 (5 Members)</td>
<td>Set 5 (5 Members)</td>
<td>Disgust</td>
<td>GTTS &gt; STTS</td>
</tr>
</tbody>
</table>

GTTS : Generated Text to Speech  
STTS: Standard Text to Speech  
(Commercially available)

From this proposed work the performance are inferred that the speech of five speakers with respect to the quality of the synthesized speech and variations in natural speech related to prosody has been enhanced. For developing an emotion based TTS output using prosody features we have taken different sets of samples with diversified emotions. For performance measure we have taken 5 sets of GTTS and STTS with different emotions. From the performance measure we inferred that GTTS has produced a high-quality Tamil text-to-speech system. Based on part-of-speech analysis, prosodic modeling and non-uniform units Tamil text into natural speech has been produced[15][16]. These technologies efficiently enhance thenaturalness and the precious quality of the TTS system.
REFERENCES


