Concatenation of syllable by anchor
frame to improve Naturalness in speech
synthesis for Marathi language (India)

Pravin M Ghate
National Institute of Electronics & Information Technology
Dr.Babasaheb Ambedkar .Marathwada, University.
Aurangabad, India
pmghate@gmail.com

Suresh Shirbahadurkar
Zeal College of Engineering & Research Narhe Pune,
Savitribai Phule Pune University,
Pune India
shirsd@gmail.com

May 26, 2018

Abstract

Speech synthesis is the method to convert text in to
speech waveform. Current state-of-the-art text-to-speech
systems for Indian language produce intelligible speech but
lack the prosody of natural utterances using concatenation
speech synthesis. It can be useful for various application,
car navigation, announcements in railway stations, response
services in telecommunications, and e-mail reading. It has
been observed that widely used approach for speech syn-
thesis is based on concatenation of segment, this approach
is called as concatenation technique. This method uses
the pre-recorded unit of speech, which preserve the nat-
urality and intelligibility of the Marathi language. The
quality of the synthetic speech is the direct function of the
available syllable. In context of Indian languages, syllable units are found to be much better choice than units like phone, diphone and half phone [1]. For good quality synthesis, syllable all units of the language should be present and join of syllable. The issues not addressed in the previous works on syllable based synthesizer for Indian languages is concatenation of the syllable. This thesis addresses two problems in speech synthesis, one is to improve the naturalness of synthetic speech in syllable-based speech synthesis. The other is the methods of improving the quality syllable for the concatenating, in order to achieve more flexible speech synthesis. To deal with the former problem, we focus on two factors: (1) an algorithm for segmentation of speech into syllable as per the linguistic rule and (2) Concatenation of syllable. In this paper, a novel method for improving the quality of syllable for Marathi language, is proposed. This method is used to generate a combination of two speech units like syllables for generation better synthetic sound in terms of naturalness and intelligibility [1]. Experimental work carried out on speech data base for Marathi language collected from Language Technologies Research Center (LTRC), which is one of the center under International Institute of Information Technology (IIIT) Hyderabad the performance of the proposed algorithm has been evaluated with state-of-the-art technique like MOS (Mean Opinion Scores) Test (Ref in: The International Telecommunication Union, ITU-T Recommendation P.85, http://www.itu.int/rec/T-REC-P.85-199406-I/en, 1994.) Mean Opinion Scores (MOS) is the test to check the quality of synthetic speech [3], we have also proposed segmentation algorithm for speech signal. Performances of the proposed algorithms have been evaluated on the parameters such as number of sample point in speech syllable, Pitch, Energy of speech and frame of the speech signal [4]. Result of proposed algorithm are encouraging due to the selection of suitable features as compared to the previous methods. Analysis has been carried out for the proposed algorithms on database of syllable which is generated form word sound. This research work aims to implement concatenation approach (using syllable as speech units) for reduction of database size and to prepare database of for syllables
so that they can be reused for synthesizing new words.

Key Words: Concatenation, Marathi Phonology, Speech Synthesis, syllable, TTS,

1 Introduction

Speech processing technology has been a mainstream area of research for more than 50 years. The ultimate goal of speech research is to build systems that mimic human capabilities in understanding, generating and coding speech for a range of human-to-human and human-to-machine interactions. Speech Synthesis is the technique of artificially generating human speech [1]. A system doing this synthesis function is called as synthesizer. A text to speech system normal text of any language for which it is design into speech. It is widely used in audio reading devices for blind people now days. The basic block shown in the Figure: 1.1

![Figure: 1.1 Basic block diagram of Speech synthesis](image)

2 System development

In the last few years however, the use of text-to-speech conversion technology has grown far beyond the disabled community to become a major adjunct to the rapidly growing use of digital voice storage for voice mail and voice response system. The system development shown in Figure 1.2

Proposed block diagram for System development
The process of converting the sentence into syllable is called segmentation. This segmentation is performed according to the linguistic rule of Marathi language. The system implements two major algorithms:

1. Segmentation algorithm
2. Syllable concatenation algorithm

The system has two inputs, like speech signal and text. These two inputs combine according to the algorithm and consideration of parameters. The accepted speech signal and text are processed further for storing in the data corpus. Mapping of text with the sound unit stored in the pre-processing is carried out. Figure 1.2 shows the complete system development work in three stages, in the first stage, it accepts the speech sentence in Marathi language on the same stage form other side it accept the Marathi text. In the last stage speech is generated. The outcome of whole algorithm is to generate the syllable and concatenation of the syllable with signal processing. Details experimentation of these algorithm is covered in the following section.

3 Algorithm syllable concatenation to form the word

This algorithm is for the concatenation of syllable naturalness and intelligibility. Our earlier work on speech synthesis has shown that conversation of speech signal into the syllable [5]. Nevertheless, audible artifacts are present due to discontinuities in pitch, energy,
and formant trajectories at the joining point of the units. There are two algorithm propose for the syllable concatenation for improving the quality of syllable for Marathi language.

1. Concatenation of syllable using the number of sample
2. Concatenation of syllable using the Pitch of Frame at joint

The first algorithm for concatenation the two syllable using the number of samples.

Step 1: Take the two syllable from database (Lets Consider the S1 first syllable and S2 second syllable)
Step 2: Find the length of syllables (S1 and S2)
Step 3: Take the last 1000 sample of first syllable (S1)
Step 4: Take the first 1000 sample of second syllable (S2)
Step 5: Find the mean of selected point
Step 6: Replace mean point at joint point
Step 7: Generate the speech after concatenation

In this paper, we present some minimal signal modification techniques for reducing these artifacts. After the segmentation of sentence into syllable. These syllable is the stored in the in data corpus of syllable. It is extremely tough to make a machine which sounds identical to human. Hence the best text to speech (TTS) algorithm ever made sounds robotic, unless and until human speech itself is involved in it. But it is not possible to create a database of each and every word possible in any language. Syllable based concatenative speech synthesis (CSS) leads to the formation of new words from existing words and syllables in the database. The most important qualities of a speech synthesis system are naturalness and intelligibility. The present Marathi TTS system is capable of automatically producing speech by storing small segments of speech and concatenating them when required. The objective of this algorithm is to increase naturalness and intelligibility at the joint point.

Flow chart for concatenation of two syllable using smoothing and filtering
In this both cases, minimum changes suggested in the original signal. Among different methods of speech improvement, pitch modification is one of the simplest methods. If the pitch is modified at the concatenation joint, the glitch or spectral mismatch can be reduced [6]. This chapter explains how syllable concatenating techniques are implemented for more naturalness of resulting speech output. The aim of this work is to increase the naturalness of text to speech synthesis by comparing pre-recorded speech (original speech) with concatenation algorithm processed output and improving the Speech quality. The system produces synthetic speech with desired pitch scale and time scale modification for arbitrary input speech signal so that there is no spectral mismatch at the point of concatenation. The second algorithm is used for concatenation of the syllable using the pitch parameter from different .The synthesized speech quality was clearly natural sounding, but there were audible artifacts causing a drop in the perceived overall quality. In this work we have attempted to understand the causes for these artifacts and present methods for improving the quality of the synthesized speech by applying minimal prosodic modifications.
4 Algorithm Concatenation of two syllable using Anchor Frame

Step 1: Take the two syllable from database (Lets Consider S1: first Syllable, S2: second Syllable)

Step 2: Convert the syllable into the Frames

Step 3: Take the Frame containing highest last two Pitch of first syllable

Step 4: Take the Frame containing highest first two Pitch of second syllable

Step 5: Interpolation of both the syllable (S1 and S2)

Step 6: Stored the values of syllable (S1 and S2) into Matrix (Consider the values of M1 First Matrix and M2 Second Matrix)

Step 7: Check the size of the Matrix & normalize the size Matrix

Step 8: Take the average of the two matrix

Step 9: Generate the new frame form the average of Matrix called Anchor Frame

Step 10: Insert the anchor frame between the two syllable (S1 and S2) for Smoothing Second algorithm for concatenation of two syllable using Anchor Frame

Figure: 1.4 Second algorithm for concatenation of two syllable using Anchor Frame
For generation of the word the syllable is the unit used for concatenation. Further the syllable is divided into the frames of 10 to 15 msec. The two syllable from database consisting the many frames. Out of the many frame consider the last frame of the first syllable and first frame of second syllable. Store the vales of maximum two pitch of the two frame and do the interpolation. The new frame generate from two frames is called as anchor frames.

5 Experimental results concatenation of syllable with smoothing and filtering signal

The objective of this research smooth concatenation of the syllable to generate the quality synthesized speech. Concatenation of two syllable using smoothing and filtering after experimentation has been carried out. It is suggested that take the last hundred (1000) point from first syllable and first hundred (1000) point from second syllable. Take the mean of these sample and insert the mean values in between the two syllable which is selected from the database.

5.1 Filtering and windowing

For joint of two syllable first we take the mean of using additive smoothing. Then hanning window we find the point. Here at the output take the syllable from speech corpus and join according to concatenation algorithm. The following equation generates the coefficients of a Hanning window

\[ w(n) = 0.5(1 - \cos(2\pi n/N)), 0 \leq n \leq N \]  

The window length \( L = N + 1 \).

Here the values for \( N=100 \) point. End 100 point from syllable 1 and start 100 point from syllable 2. The reason for selecting 100 point is we get the better quality of output speech. Let us consider the two syllable S1 and S2. The novel work in this research is to propose concatenation algorithm. If we join the two syllable the naturalness of the signal is increased

Example: 1
Concatenation of syllable (Syllable-1) + (Syllable -2) For experiment to the second algorithm we have taken the syllable from speech corpus which is processed i.e. and concatenation done after filtering and smoothing. one of example which is shown in the figure 1.5 We have two words and , extract the and and form the new word.

Example 1

Figure 1.5 : Concatenation of Syllable

Figure: 1.6 (a) Concatenation of syllables before processing, (b) X-axis number of sample (y) amplitude of the signal
In many experiments in science, the true signal amplitudes (y-axis values) change rather smoothly as a function of the x-axis values, whereas many kinds of noise are seen as rapid, random changes in amplitude from point to point within the signal. In some cases to attempt to reduce the noise by a process called smoothing. In smoothing, the data points of a signal are modified so that individual points that are higher than the immediately adjacent points (presumably because of noise) are reduced, and points that are lower than the adjacent points.

Table: 1.1 Average 10 point showing smoothing of joint at syllable

<table>
<thead>
<tr>
<th>Sr.No</th>
<th>Before Smoothing</th>
<th>After Smoothing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0023</td>
<td>0.0018</td>
</tr>
<tr>
<td>2</td>
<td>0.0024</td>
<td>0.0011</td>
</tr>
<tr>
<td>3</td>
<td>0.0058</td>
<td>0.0005</td>
</tr>
<tr>
<td>4</td>
<td>0.0042</td>
<td>-0.0001</td>
</tr>
<tr>
<td>5</td>
<td>0.0050</td>
<td>-0.0004</td>
</tr>
<tr>
<td>6</td>
<td>0.0039</td>
<td>0.0006</td>
</tr>
<tr>
<td>7</td>
<td>0.0010</td>
<td>0.0000</td>
</tr>
<tr>
<td>8</td>
<td>0.0058</td>
<td>0.0001</td>
</tr>
<tr>
<td>9</td>
<td>0.0020</td>
<td>0.0040</td>
</tr>
<tr>
<td>10</td>
<td>0.0009</td>
<td>0.0049</td>
</tr>
</tbody>
</table>

6 Experimental result of anchor frame concatenation algorithm for syllable.

The second concatenation syllable anchor algorithm is used to join to syllable. In this method anchor frame is generated form the two syllable joint. The last segment of first syllable containing the highest two pitch and first segment of last second syllable containing the two pitch. All the values stored in the matrix and create the new frame that is called as the anchor frame. This new frame is used to join the syllable smoothly. Following are the some example for concatenation the syllable using the anchor frame.

Table 1.2 some of the example of syllable concatenation
Following diagram show the results of anchor frame concatenation algorithm for syllable, As per the flow chart process every stage shows the waveform.

**Example 1**

The Figure 1.8 shows waveform of syllable taken from the database for concatenation of syllable to form the word. In this diagram x axis shows amplitude of the signal and y-axis shows number of sample. The first syllable and second syllable shows the waveform in this diagram. The figure 1.9 shown theframes of speech signal in the different colours for further processing.
To remove the discontinuity at the joint of syllable and to maintain smooth effect in output speech, interpolation of signal carried out on two frame. The process of interpolation is shown in Figure 1.10. The figure shows before interpolation and after interpolation of the last frame of first syllable and first frame of second syllable.

The process of interpolation increase the number of sample in the both the frame. By selecting the last two pitch of first syllable and first two pitch of second syllable, the new frame is formed. This frame is called the anchor frame. The anchor frame shows in the Figure 1.11

This anchor frame is adjusted in the between the two syllable so that new word formed which shows in Figure 1.12
The GUI (Graphical user interface) has been designed for concatenation of syllables, as shown in Figure 1.13, for evaluating the quality of synthesized speech. The GUI consists of two small windows; these two small windows are used for accepting the input for concatenation and the play button is provided for listing the sound. As the experiment is related to speech synthesis, the window accepts the text, accepts the text map with the syllable database, and text converted into speech. Now the second task is to concatenate the syllable and the further formation of the word. In the same GUI, another window shows the energy of each syllable and boundaries of the syllable.

For example, syllable 1 syllable 2. These two syllable text is inputted to the window for concatenation of two syllable.

Figure 1.12. Inserting the anchor frame in Syllable

7 Testing Methodologies for Intelligibility and Naturalness
Some of the examples for concatenation the syllable. Using the anchor frame concatenation algorithm

Table 1.3: list of syllable for formation of the word

<table>
<thead>
<tr>
<th>Sr.No</th>
<th>Syllable-1</th>
<th>Syllable-2</th>
<th>Formation of the word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>अ</td>
<td>अ</td>
<td>पात्र</td>
</tr>
<tr>
<td>2</td>
<td>ए</td>
<td>ए</td>
<td>पात्र</td>
</tr>
<tr>
<td>3</td>
<td>ए</td>
<td>ए</td>
<td>पात्र</td>
</tr>
<tr>
<td>4</td>
<td>ए</td>
<td>ए</td>
<td>पात्र</td>
</tr>
<tr>
<td>5</td>
<td>ए</td>
<td>ए</td>
<td>पात्र</td>
</tr>
<tr>
<td>6</td>
<td>ए</td>
<td>ए</td>
<td>पात्र</td>
</tr>
<tr>
<td>7</td>
<td>ए</td>
<td>ए</td>
<td>पात्र</td>
</tr>
<tr>
<td>8</td>
<td>ए</td>
<td>ए</td>
<td>पात्र</td>
</tr>
<tr>
<td>9</td>
<td>ए</td>
<td>ए</td>
<td>पात्र</td>
</tr>
<tr>
<td>10</td>
<td>ए</td>
<td>ए</td>
<td>पात्र</td>
</tr>
</tbody>
</table>

8 Conclusion:

The paper explains the two algorithms for concatenation of the syllable. The first algorithm proposes the maximum and minimum point. In this algorithm, the 1000 point of the last segment of the first syllable and the 1000 point of the first segment of the second syllable. On the joint point, signal processing is carried out for smoothing the signal. Still, results are not up to the mark. So we propose the second algorithm, in which the syllable is divided into a number of frames of 12 msec. There are two syllables needed for combining. Take these two syllables from the database. Take the last frame of the first segment and the first frame of the second syllable. Using the point from two frames, create a new anchor frame.

Acknowledgment

Author would like to thank management of JSPMs Rajarshi Shah College of Engineering, Tathawade, for continuous motivation and support.

References


