Research Review on Text-to-Speech Systems and Speech Synthesizing Techniques

Gargi B. Shah¹ and Dr. Priti S. Sajja²
¹(Assistant Professor, Department of Computer Science, VNSGU, Surat, India)
²(Professor, Department of Computer Science, SPU, V.V. Nagar, India)

Abstract: The most common, easy and feasible ways of communication is speech. Due to wide range of applications and physically disabled people requires an application to convert text to speech. Now a days Gujarati language is also widely used as a communication medium. Many systems has been developed so far to convert text to speech with one or more limitations. This review paper includes overview about text to speech systems developed for Gujarati language. The techniques to develop speech synthesizer are included. All individual technique is having limitations. So, to overcome these limitations, methodology is provided to develop a hybrid synthesizer for text to speech conversion for Gujarati Language. As a result, A good speech synthesizer achieves two quality measures of text to speech system: Naturalness and Intelligibility. The produced speech than pass through Super position algorithm developed using Formant Synthesizer and Time-domain Synthesizer to achieve more prosody on output and synthesized speech.

Keywords: Text to Speech, Artificial Neural Network, Grapheme to Phoneme, Formant Synthesizer, Time-Domain Synthesizer

I. INTRODUCTION

Speech is generally used as way of communication in our daily life. In the era of information and communication technology, the communication made by digital ways is important in speech communication research and application. The text-to-speech (TTS) conversion system will play very important role for physically disabled people with visually impairment, illiterate masses and wide range of digital speech based applications.

TTS system are in great demand for the Indian languages specially for Gujarati language, as very less work has been done for Gujarati language.

The speech synthesizers are used by TTS systems to convert text in to speech / voice. It is nothing but digitally developed human speech. A computer system used for developing such artificial human speech is known as speech synthesizer. The TTS system works on two basic properties: (1) Text Processing (2) Speech Synthesis Techniques.

The little amount of work has been carried out on Gujarati language and is converted in to speech. This review paper describes already developed TTS systems for Gujarati language, such as Dhvani, Shruti, HP labs India TTS system, Vani, SAFA, eSpeak etc. It also includes different speech synthesizer techniques, which can be combined to develop hybrid synthesizer. The comparisons between existing TTS systems and synthesizer techniques give a better way to develop a TTS system.

Problems with Existing System

All Indian language TTS systems have common phonetic base. At present, very few TTS systems are in place for Gujarati language.

The problems with current TTS systems are :

1) They are with inadequate prosodic models.
2) The following quality measures for speech synthesizer are not greatly achieved.
   (A) Naturalness: Similarity to natural human speech
   (B) Intelligibility: Ease of understanding by the listener.
3) They does not support for word document, PDF or scanned data.
4) All TTS system are developed using single synthesizer, which is having limitation in one or the other way.
5) The currently available TTS systems reads single digit number, they do not read number as a whole.

II. RELATED WORK

The text-to-speech (TTS) is greatly in demand, which can help illiterate mass and visually disabled people. Here, some of the TTS systems for Indian languages are described:

1) Dhvani – (for Indian languages)
Dhvani system was awarded by FOSS India award in the year 2008.

Characteristics:
(A) This system was designed for Indian languages, developed by Dr. Ramesh Hariharan, trustee of Simputer trust at Institute of Science, Bangalore in the year 2000.
(B) Languages like Hindi, Malayalam, Bengali, Oriya, Punjabi, Gujarati, Telugu and Marathi are supported by this system and separate module for every language is developed.
(C) This system uses diaphone Concatenation algorithm.
(D) All sound files are stored in the database as ‘gsm’ compressed files.
(E) This system works on GNU/Linux platforms.
(F) Each system requires Unicode parser.
(G) It stores data using Syllable database.

This system does not provide prosody on the output, simply sound units at pitch periods are concatenated and plays them out. The database size currently is only around 1 MB. [7]

(2) Shruti – (An Embedded text-to-speech (TTS) system)

A Mukhopadhyay developed this system at Indian Institute of Technology, Kharagpur in the year 2006.

Characteristics:
(A) Concatenative Speech Synthesis technique was used to develop this TTS system.
(B) This system was built specifically for Bengali and Hindi languages.
(C) This system can be extended to other Indian languages as well.
(D) The system can be used or ported to Compaq iPaq and the Casio Cassiopeia processor families.
(E) It uses Partneme Database.

Shruti text-to-speech (TTS) system does not support Gujarati language. [7]

(3) TTS system by HP Labs India

The TTS system developed by HP labs India is based on open source TTS framework. A generic Grapheme in to phoneme (G2P) conversion system has been developed at HP labs India. It is a language Independent engine, which requires language dependent information in the form of lexicon, rules and mapping. [7]

(4) Vani – An Indian Language Text to Speech Synthesizer

This system is developed at IIT Bombay, India. All TTS systems specify users to what is to be spoken, but it does not specify how it is to be spoken.

The vTrans encoding scheme has been included in Vani system. A vTrans help the person to encode the text to be spoken and also helps how it is to be spoken. A Signal processing module of this system embeds variations to the sound database to be incorporated in speech.

This TTS system uses Concatenative Synthesis technique, and it supports for Hindi Language. It includes parameters to control speech like pitch, volume and duration is given by user. [7]

(5) SAFA (Screen Access for All)

SAFA is not actually a speech synthesizer. It is a program developed by National association for the Blind, New Delhi. It is a screen reader. This system works by providing a platform to detect text language and accordingly calls TTS for speaking it. The latest version of SAFA supports following language: Hindi, English, Sanskrit, Tamil, Marathi, Bengali, Nepali, Gujarati, Kannada and Telugu. [7]

(6) eSpeak

eSpeak is compact, multi-language , open source text-to-speech synthesizer. This version is a SAPI5 compatible windows speech engine which should work with screen readers such as Jaws, NVDA and Windows-Eyes. There is also a version of eSpeak which can run as a command-line program. [7]

**Table1: Comparison of TTS Systems for Indian Languages**

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Name of the System</th>
<th>Synthesis Strategy</th>
<th>Language Support</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dhvani</td>
<td>Diphone Concatenation Synthesis</td>
<td>Hindi, Malayalam, Bengali, Oriya, Punjabi, Gujarati, Telgu, Marathi,</td>
<td>No prosody on output.</td>
</tr>
<tr>
<td>2</td>
<td>Shruti</td>
<td>Concatenative Synthesis</td>
<td>Bengali, Hindi</td>
<td>It does not support Gujarati Language</td>
</tr>
<tr>
<td>3</td>
<td>Text analysis module created by HP lab India.</td>
<td>Concatenative Synthesis</td>
<td>Hindi</td>
<td>No prosody for specific language</td>
</tr>
<tr>
<td>4</td>
<td>Vani</td>
<td>Concatenative Synthesis</td>
<td>Hindi</td>
<td>No support for Gujarati Language</td>
</tr>
</tbody>
</table>
From the above comparison, Only Dhvani TTS system supports for Gujarati language. But no prosody rules are applied on the output.

III. EXPECTED METHODOLOGY

The process for converting text-to-speech is divided in to number of stages. The architecture of text-to-speech (TTS) system is divide in to two main properties.

(1) Text Processing
(2) Speech Synthesis Technologies

![Figure 1: Modules of Text-to-Speech Conversion Methodology](image)

Module 1: Text Processing

Sub processes of Text Processing Module:
(A) Text Normalization
(B) Dictionaries
(C) Text to Phoneme
(D) Prosody Rules

(1) Text Normalization

The raw text contains numbers and abbreviations. This sub process converts them in to written words. It searches for numbers, times, dates and other symbolic representations.

(2) Dictionaries

Pronunciations of some words are different than they are recorded in the dictionary. The exception dictionary is maintained to store the phonemic transcription of these words.

![Figure 2: Block Diagram for Text Processing Module](image)

(3) Text to phoneme

Pronunciation of each word according to its spelling is known as text to phoneme process. Following two approaches are used for text to phoneme conversion.

(A) Dictionary based Approach: In this approach, All the words of a target language is stored in database with correct pronunciation. While converting text-to-speech, the word is first looked up in to the dictionary and word will be replaced with pronunciation of the word. The advantage of this approach is that it is quick and accurate, but it fails if the word is not found in the dictionary.
(B) **Rule based Approach:** Every word is having pronunciation rule based on spelling of the word. These pronunciation rules are determined by this approach and applied on the words while converting it in to speech. This approach works for any inputted word, but rule complexity grows for irregular and spelling and pronunciation.

(4) **Prosody Rules**

The most challenging problem is to find out intonation, stress and duration from written text. The intonation implies how the pitch pattern or fundamental frequency changes during speech. These features together are called prosodic features and considered as the melody, rhythm, and emphasis of the speech at the perceptual level. [7]

**Module 2: Speech Synthesis Technologies**

Artificially generation of human speech from text is known as Speech Synthesis Technologies. The word synthesis refers to putting together parts of elements so as to form a whole. The computer system used for speech synthesis is known as speech Synthesizer. Speech synthesizer can be hardware based or software based.

The required database in the conversion of character to sound is recorded by “PRAAT” software and saved in .wav format in the directory. Concatenation technique will be used for synthesis process and “MATLAB” software will be used.

The quality of speech synthesizer is measured in two primary factors:
1. Naturalness: similarity to normal human speech.
2. Intelligibility: Ease of understanding by the user.

There are two approaches for measuring speech quality:
1. Subjective Approach: In this approach, hear the recorded speech of a processed voice file and measure the speech quality based on opinion scale and MOS (Mean Opinion Score) is calculated.
2. Objective Approach: In this approach, machine based automatic assessment is done. The measure is derived from inputted text and natural speech corpus and is inversely propositional to overall speech quality. The higher the number of concatenation, lower is the quality.

**Methods of Speech Synthesis:**

The speech synthesis methods are usually classified in to three categories:
1. Articulatory Synthesis
2. Formant Synthesis.
3. Concatenative Synthesis

**Articulatory Synthesis**

Human vocal organ are modelled perfectly with an Articulatory synthesis. So, this is satisfied method to produce high-quality synthetic speech. With this method, the computational load is very higher as compared to other methods. So, It is most difficult method to implement. Thus, it has received less attention than other synthesis methods and has not yet achieved the same level of success.

Articulatory synthesis typically involves models of the human articulators and vocal cords. The articulators are usually modelled with a set of area functions between glottis and mouth. For rule-based synthesis the articulatory control parameters are lip aperture, lip protrusion, tongue tip height, tongue tip position, tongue height, tongue position and velic aperture. Phonatory or excitation parameters may be glottal aperture, cord tension, and lung pressure.

The vocal tract muscles cause articulators to change and move the shape of vocal tract to generate different sounds. The X-ray analysis of natural speech is taken as an input to collect data for articulatory model. However, data collected by X-ray analysis are usually only 2-D, while the real vocal tract is naturally 3-D, so the rule-based articulatory synthesis is very difficult to optimize due to the unavailability of sufficient data of the motions of the articulators during speech. Other deficiency with articulatory synthesis is that X-ray data do not characterize the masses or degrees of freedom of the articulators. Also, the movements of tongue are so complicated that it is almost impossible to model them precisely. [2][3]

**Formant Synthesis**

The source-filter model based Formant Synthesis method is widely used synthesis method during last decades. There are two basic structures of these methods. They are parallel and cascade, but for better performance some kind of combination of these is usually used. Formant synthesis also provides infinite number of sounds which makes it more flexible than for example concatenation methods.

At least three formants are generally required to produce intelligible speech and up to five formants to produce high quality speech. Each formant is usually modelled with a two-pole resonator which enables both pole-pair formant frequency and its bandwidth to be specified.

A cascade formant synthesizer consists of band-pass resonators connected in series. The input of each following resonator is the output of leading formant resonator. The cascade structure needs only formant frequencies as control information. In cascade structure the relative formant amplitudes for vowels do not need individual controls. This is main advantage of cascade structure.
The non-nasal voiced sound found better for cascade structure because it needs less control information than parallel structure. So, it is simpler and easier to implement. However, with cascade model the generation of fricatives and plosive bursts is a problem.

A parallel formant synthesizer consists of resonators connected in parallel. Sometimes extra resonators for nasals are used. The input to all formants is excitation signal, which are passed simultaneously to each formants and their outputs are summed. Adjacent outputs of formant resonators must be summed in opposite phase to avoid unwanted zeros or ant resonances in the frequency response. The parallel structure enables controlling of bandwidth and gains for each formant individually and thus needs also more control information.

The parallel structure has been found to be better for nasals, fricatives, and stop consonants, but some vowels cannot be modelled with parallel formant synthesizer as well as with the cascade one.[2][3]

**Concatenative Synthesis Method**

In this Cut and Paste method, waveforms are stored in database. Some joining rules are applied to these segments to produce sound. To achieve intangibility and naturalness in sound, connecting pre-recorded natural utterances is probably the easiest way. Only one speaker and one voice can participate in Concatenative synthesizer. To find out correct unit length is the main aspect of this method. The memory requirement will be increased with longer unit length, but with longer units high naturalness, less concatenation points and good control of coarticulation are achieved. With shorter unit, less memory is needed, but the sample collecting and labelling procedures become more difficult and complex.

**Units for Concatenative Synthesis:**

a) **Phone:** Speech is sequence of sounds, while Phone is single unit of sound.

b) **Diphone:** A diphone is defined as the signal from either midpoint of a phone or point of least change within the phone to the similar point in the next phone.

c) **Triphone:** A triphone is a section of the signal taking in a sequence going from middle of a phone completely through the next one to the middle of the third.

There are several problems in concatenative synthesis compared to other methods.

1. Distortion from discontinuities in concatenation points.
2. For long concatenation units, such as syllables and words, Memory requirements are usually very high,
3. Data collecting and labelling of speech samples is usually time-consuming. [2][3]

**Table 2: Comparison between Different Speech Synthesis Techniques**

<table>
<thead>
<tr>
<th>Articulatory Synthesis</th>
<th>Formant Synthesis</th>
<th>Concatenative Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Data are derived from X-ray analysis of natural speech. X-ray data are usually in 2-D, when vocal tract are normally 3-D. So, it is difficult to</td>
<td>- Some vowels cannot be modelled with parallel formant synthesizer as well as with the cascade one.</td>
<td>- Distortion from discontinuities in concatenation points.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Memory requirements are usually very high, especially when long units are used.</td>
</tr>
</tbody>
</table>
optimize due to unavailability of sufficient data of the motion of the articulators during speech.

- X-ray data do not characterize the masses or degree of freedom of the articulator.
- The moments of tongue are so complicated that it is almost impossible to model them precisely. It is based on Genioglossus, Styloglossus and Hyoglossus muscles of the tongue. They are responsible for the main displacement and shaping of the overall tongue structure.

From the above study, there is no single method which is best suitable for TTS system, but combination of previously described methods may create hybrid synthesizer. Synthesized speech can also be manipulated afterwards with normal speech processing algorithms, such as we can add echo to improve the speech. But with this approach computational load of the system may get increased.

To overcome the above mentioned limitations of individual synthesis methods, the combination of synthesis methods results in hybrid synthesizer.

Time domain synthesis and Formant synthesis approaches combinedly generates hybrid synthesizer, with high quality speech segment derived from time domain synthesis and homogeneous speech allows good control of fundamental frequency derived from formant synthesis.

**Phonemic and Prosodic Information**

![Figure 5: Basic Idea of Hybrid Synthesis System](image)

Several methods and techniques for determining the control parameters for a synthesizer may be used. Recently, the artificial intelligence based methods, such as Artificial Neural Networks (ANN), have been used to control synthesis parameters, such as duration, gain, and fundamental frequency. Neural network use a set of processing elements or nodes analogous to neurons in the brain. These processing elements are interconnected in a network that can identify patterns in data as it is exposed to the data.

**IV. REFERENCES**


http://indusedu.org

This work is licensed under a Creative Commons Attribution 4.0 International License


