IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# Devanagri Text To Speech Conversion

**Poonam S.Shetake**
Bharati Vidyapeeth's College of Engineering, Kolhapur, India
poonamshetake3@gmail.com

## Abstract

Language is the ability to express one's thoughts by means of a set of signs, whether graphical gestural, acoustic, or even musical. It is distinctive nature of human beings, who are the only creatures to use such a structured system. Speech is one of its main components. It is by far the oldest means of communication between human being and also the most widely used. No wonder, then, that people have extensively studied it and often tried to build machines to handle it in acoustic way. Most of the Information in digital world is accessible to a few who can read or understand a particular language. Language technologies can provide solutions in the form of natural interfaces so the digital content can reach to the masses and facilitate the exchange of information across different people speaking different languages. These technologies play a crucial role in multi-lingual societies such as India which has about 1652 dialects/native languages.

## Introduction

Language is the ability to express one's thoughts by means of a set of signs, whether graphical gestural, acoustic, or even musical. It is distinctive nature of human beings, who are the only creatures to use such a structured system. Speech is one of its main components. It is by far the oldest means of communication between human being and also the most widely used. No wonder, then, that people have extensively studied it and often tried to build machines to handle it in acoustic way. Most of the Information in digital world is accessible to a few who can read or understand a particular language. Language technologies can provide solutions in the form of natural interfaces so the digital content can reach to the masses and facilitate the exchange of information across different people speaking different languages. These technologies play a crucial role in multi-lingual societies such as India which has about 1652 dialects/native languages.

A text to speech converter converts normal language text into speech. 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 diphones 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.

## Present Theories And Practices

Undoubtedly, ability to speak is the most important way for humans to communicate between each other. Speech conveys various kind of information, which is essentially the meaning of information speaking person wants to impart, individual information representing speech and also some emotional filling. Speech production begins with the initial formalization of the idea which speech wants to impart to the listener. Then speech converts this idea into the appropriate order of words and phrases according to the language. Finally, his brain produces motor nerve commands, which move the vocal organs in an appropriate way. Understanding of how human produce sounds forms the basis of speech identification.

The sound is an acoustic pressure formed of compressions and rarefactions of air molecules that originate from movements of human anatomical structures. Most important components of the human speech production system are the lungs (source of air during speech), trachea (windpipe), larynx or its most important part vocal cords (organ of voice production), nasal cavity (nose), soft palate or velum (allows passage of air through the nasal cavity), hard palate (enables consonant articulation), tongue, teeth and lips. All these components, called articulators by speech scientists, move to different positions to produce various sounds. Based on their production, speech sounds can also be divided into consonants and voiced and unvoiced vowels.

From the technical point of view, it is more useful to think about speech production system in terms of an acoustic filtering operation that affects the air going

from the lungs. There are three main cavities that comprise the main acoustic filter. According to they are nasal, oral and pharyngeal cavities. The articulators are responsible for changing the properties of the system and form its output. Combination of these cavities and articulators is called vocal tract.

In a speech recognition task, we are interested in the physical properties of human vocal tract. In general it is assumed that vocal tract carries most of the speech related information. However, all parts of human vocal tract described above can serve as speech dependent characteristics. Starting from the size and power of lungs, length and flexibility of trachea and ending by the size, shape and other physical characteristics of tongue, teeth and lips. Such characteristics are called physical distinguishing factors. Another aspect of speech production that could be useful in discriminating between speeches is called learned factors, which include speaking rate, dialect, and prosodic effects.

The most important qualities of a speech synthesis system are "naturalness" and "*Intelligibility*". Naturalness describes how closely the output sounds like human speech.

Intelligibility is the ease with which the output is understood. The ideal TTS system is both of the above. There are two primary technologies for TTS Systems:

a) Concatenative synthesis.
b) Formant synthesis

**Concatenative Synthesis :** In this approach Synthesis is done by natural speech. This methodology has the advantage in its simplicity, i.e. there is no mathematical model involved . Speech is produced out of natural, human speech. Concatenative synthesis is based on the concatenation of segments of recorded speech. Generally, concatenative speech produces the most natural sounding synthesized speech output. It has three sub-types:

1. Unit selection synthesis:

It uses large databases of recorded speech. During database creation, each recorded utterance is segmented into phones, diphones etc.

2. Diphone synthesis:

It uses a minimal speech database containing all the diphones occurring in a language. The number of diphones depends on the phonotactics of the language.

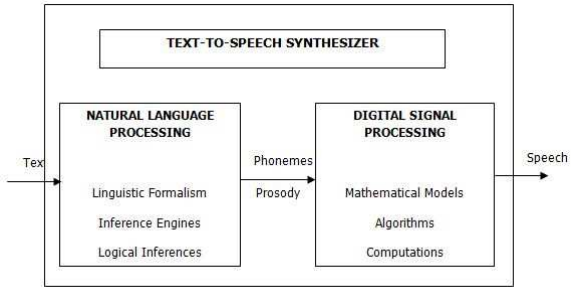
3. Domain specific synthesis:

It concatenates pre-recorded words and phrases to create complete utterance. It is used in applications where the variety of texts the system will output is limited to a particular domain, like weather reports etc. It is simple to implement and has been in commercial use for a long time.

**Formant Synthesis:** Formant synthesis does not use human speech samples at runtime. Instead, the synthesized speech output is created using an acoustic model. Parameters such as fundamental frequency , voicing , and noise levels are varied over time to create a waveform of artificial speech. This method is sometimes called rules-based synthesis; however, many concatenative systems also have rules-based components.

Many systems based on formant synthesis technology generate artificial, robotic-ech that would never be mistaken for human speech. However, maximum naturalness is not always the goal of a speech synthesis system, and formant synthesis systems have advantages over concatenative systems. Formant-synthesized speech can be reliably intelligible, even at very high speeds, avoiding the acoustic glitches that commonly plague concatenative systems. High-speed synthesized speech is used by the visually impaired to quickly navigate computers using a screen readers. Formant synthesizers are usually smaller programs than concatenative systems because they do not have a database of speech samples. They can therefore be used in embedded systems, where memory and microprocessor power are especially limited. Because formant-based systems have complete control of all aspects of the output speech, a wide variety of prosodies and intonations can be output, conveying not just questions and statements, but a variety of emotions and tones of voice.

## System Block Diagram




## Proposed Work

The block diagram shows two major blocks of the system: natural language processing and the digital signal processing (mathematical models)

The text is given as the input to the system. The text is compared with the database stored in the system. The database includes the wave files as well as the script. When the text is received as the input, it is converted into phonemes and prosody.

In human language, a phoneme is the smallest posited structural unit that distinguishes meaning. Phonemes

are not the physical segments themselves, but, in theoretical terms, cognitive abstractions or categorizations of them. An example of a phoneme is the /t/ sound in the words tip, stand, water, and cat. (In transcription, phonemes are placed between slashes, as here.) These instances of /t/ are considered to fall under the same sound category despite the fact that in each word they are pronounced somewhat differently. The difference may not even be audible to native speakers, or the audible differences not perceived. That is, a phoneme may encompass several recognizably different speech sounds, called phones. In linguistics, prosody is the rhythm, stress, and intonation of speech. Prosody may reflect the emotional state of a speaker; whether an utterance is a statement, a question, or a command; whether the speaker is being ironic or sarcastic; emphasis, contrast and focus; and other elements of language which may not be encoded by grammar.

These is transferred digital signal processing block which holds the mathematical model viz, concatenation technique. All the wave files are concatenated and a resulting file is generated which pronounces the entire text with emotions and expressions. That is the final speech is generated.

Implementation Steps:-

1. Database preparation for devnagari scripts (barakhadi)
2. Then image processing algorithm for extraction various feature from database
3. Then store feature into database as feature vectors
4. Distance matching based recognize the character
5. Once identified the char respective wave file for character is played
6. Wave file analysis fft and spectrum level to improve the characteristics for speech files
7. Final testing with joint character and testing the scheme

## References


[1] Jonathan Allen, M. Sharon Hunnicutt, Dennis Klatt, From Text to Speech: The MITalk system. Cambridge University Press: 1987. ISBN 0521306418

[2] Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. Journal of the Acoustical Society of America, 70, 321-328.

[3] P. H. Van Santen, Richard William Sproat, Joseph P. Olive, and Julia Hirschberg, Progress in Speech Synthesis. Springer: 1997. ISBN 0387947019

[4] History and Development of Speech Synthesis, Helsinki University of Technology, Retrieved on November 4, 2006

[5] http://query.nytimes.com/search/query?ppds=per&v1=GERSTMAN%2C%20LOUIS&sort=newest NY Times obituary for Louis Gerstman.

[6] Arthur C. Clarke online Biography

[7] Bell Labs: Where "HAL" First Spoke (Bell Labs Speech Synthesis website)

[8] Anthropomorphic Talking Robot Waseda-Talker Series

[9] Alan W. Black, Perfect synthesis for all of the people all of the time. IEEE TTS Workshop 2002. (http://www.cs.cmu.edu/~awb/papers/IEEE2002/allthetime/allthetime.html)

[10] John Kominek and Alan W. Black. (2003). CMU ARCTIC databases for speech synthesis. CMU-LTI-03-177. Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

[11] Julia Zhang. Language Generation and Speech Synthesis in Dialogues for Language Learning, masters thesis, http://groups.csail.mit.edu/sls/publications/2004/zhang_thesis.pdf Section 5.6 on page 54.

[12] T. Dutoit, V. Pagel, N. Pierret, F. Bataiile, O. van der Vrecken. The MBROLA Project: Towards a set of high quality speech synthesizers of use for non commercial purposes. ICSLP Proceedings, 1996.

[13] John Holmes and Wendy Holmes. Speech Synthesis and Recognition, 2nd Edition. CRC: 2001. ISBN 0748408568.

[14] Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. Speech perception without traditional speech cues. Science, 1981, 212, 947-950.