Text To Speech Synthesis for Indian Hindi Language

Bijendra Kumar
*M.Tech**
Sobhasaria group of institution

Abstract: - **Speech processing technology has been a main stream area of research for more than 50 years. The primary motivations are to provide users with a friendly vocal interface with the computer and to allow people with certain handicaps (such as blindness) to use the computer. An increasing demand in mobile and other automated based services has made speech synthesis popular in speech technology.**

A text to speech synthesis system for Hindi language is developed which synthesis the speech signal of Devnagri as well as Romanized script. The technique used is modified TD-PSOLA (Time Domain- Pitch Synchronous over lap and add) applied only at the concatenated points to remove distortion between two segments.

The concatenative technology of TD-PSOLA is well known method applied for achieving the prosody of synthesize speech. The conversion of Devnagri to Romanized script is well illustrated with the synthesis of speech signal. Here, an attempt is made with minimal errors to achieve the prosody of synthesize speech.

I. 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 (or potentially surpass) human capabilities in understanding, generating and coding speech for a range of human-to-human and human-to-machine interactions.

Speech is one of the most vital forms of communication in our everyday life. Since speech is a primary medium for communication among human beings, it is natural for the people to expect to be able to carry out spoken dialogue with computers. This involves the integration of speech technologies and language technologies. Speech synthesis is an automatic generation of artificial speech signal by the computer. In the last few years, this technology has been widely available for several languages for different platform from personal computer to stand alone system. Today, the most common interfaces for human machine interaction are still keyboards, keypads, and mice. However, an increasing necessity to interface with machines in mobile environments is leading to speech becoming a required means to interface with machines and automated information services. TTS is a complex problem that has made significant progress in the realm of concatenative systems in the last few years.With the development of new techniques such as speech synthesis and speech recognition, we are now moving into an era of more effective TTS with improved prosody in the synthesize speech. Developing a text to speech system for a language that can support inputs in other languages can be helpful to the users who know the language but are not familiar with its relative keyboard layout. Users who do not know that language at all can type in that language using their local language keyboard layout. Many times, Indian users prefer to type Hindi sentences in English Script. This fact is more prevalent over mobile SMS application, chatting and the social networking websites. Users do this because they are used to the English QWERTY keyboard and due to non-familiarity and non-availability of the Hindi keyboard.

Based on the scenario as above, the need for an effective TTS system for Hindi Language is highly felt which implemented to provide an option to the user to input the text either in Devnagri script or in Romanized English text. For such a system input string comes in form of a Romanized sentence or word (all possibilities considered) apart from the regular Hindi sentence or word, which is further processed to output a synthesized speech. The user interface also provides conversion of Hindi Text into Romanized English Text for the understanding of all those who are not familiar with the Hindi Text typed by the user. The Concatenative TD-PSOLA technique has been used to improve the prosody of synthesize speech signal.for virtually any task application. The problem of converting from text to a complete linguistic description of associated sound was one that has been studied almost as long as synthesis itself; much progress had been made in almost every aspect of the linguistic description of speech as in the acoustic generation of high quality sounds.

A TTSBOX was developed by Dutoit (2005) which performed the synthesis of Genglish (for "Generic English"), an imaginary language obtained by replacing English words by generic words.

These Vowels and Consonants with all *matra's* are separately recorded and stored. This form various phonemes of Hindi language arranged in a numerical fashion for ease of retrieving and storing. No such phonemes exist except these phonemes for Hindi language. The basis of developing a database completely relies on the Romanized script. Once the preparation of Database is complete, the process of Tokenization is done. The Tokenization links the written text with the Database.

This thesis isorganized as follows:

The next presents the Review of Literature on the previous research done in this area by various researches. NEXT describes the techniques and methods used in the

development of TTS system. NEXT presents the results obtained in the form of generation of speech from the written text by the user.The main objective of this thesis is to study the situation of today's speech synthesis technology and to focus on potential methods for future of this thesis of this thesis report. The objective of whole project is to develop good quality audio speech synthesis with a well-synchronized talking heads. Other aspects, such as naturalness, personality, platform independence, and quality assessment are also under investigation.

II. REVIEW OF LITERATURE

Speech synthesis has progressed remarkably in recent years, and it was no longer the case that state-of-the-art systems sound overtly mechanical and robotic. Before special-purpose DSP chips were introduced, synthetic speech was generated primarily on large computers, sometimes interfaced with an analogy vocal tract model. Now "speech synthesizer" devices range from inexpensive software programs for home computers to reading machines for the blind but still they all represent tradeoffs among the conflicting demands of maximizing speech quality, while minimizing memory space, algorithmic complexity, and computation time.

The quality of final synthetic speech depends on all the stages of the synthetic speech development process, neat speech editing and segmentation, accurate analysis and encoding, and complete strategy rules present better sounds. That said it is normally fairly easy to tell that it is a computer talking rather than a human, and so substantial progress is still to be made.

That used basic rules to subdivide the original problem into easier tasks which was then solved by dedicated neural networks.synthesis. In the context of HNM, speech signals were represented as a time-varying harmonic component plus a modulated noise component. The decomposition of a speech signal into these two components allows for more natural-sounding modifications of the signal (e.g., by using different and better adapted schemes to modify each component). The parametric representation of speech using HNM provides a straightforward way of smoothing discontinuities of acoustic units around concatenation points.previous system, which used a rule-based approach (Klatt model) the results were much better, even when using a limited number of parameters. From the review of literature it may be observed that the area of speech generation had been the hardest speech technology area to obtain any viable degree of success.

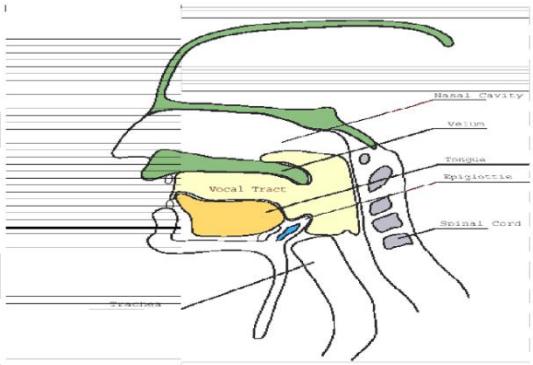
III. MATERIALS AND METHODS

This Chapter explains the materials and methods involved in the successful completion of TTS for Hindi language. The linguistic analysis and pronunciation field of English language is more complex than Hindi language. For example: In English language, the pronunciation of 'door' & 'poor' shows a vast variation. But not such happens or least occurred in case of Hindi language. Therefore preparing a TTS system for Hindi language is easier in comparison to English language.

TTS system for Hindi language is easily adaptable for two scripts. The user can write the Text either in Devnagri or Romanized script. The written Text is processed and synthesis the artificial speech.Before indulging with the approach and implementation of TTS for Hindi Language, it is must to pursue the speech production model so called as Acoustic theory of speech signal.

IV. SPEECH PRODUCTION MODEL

Synthesizing human speech is difficult due to the complexity of human speech. The production of human speech involves the lungs, the vocal folds, and the vocal tract (oral cavity, nasal cavity, and pharyngeal cavity) functioning collectively. Figure shows the organs used in speech production.



Human speech is created by an air source (lungs and the surrounding muscles) causing some type of excitation in the vocal system (vocal folds and vocal tract). The type of sound produced is determined by the vocal system's affect on the air flow. There are two types of speech produced by humans; voiced and unvoiced. With voiced speech, sound is produced from the vibration caused by air flowing through tensed vocal folds. Unvoiced speech is created from air flowing through abducted vocal folds, and the sound is produced by air flowing through a constriction the vocal tract or air being stopped and then suddenly released. Mimicking the sounds created by human speech is difficult because real continuous speech is a combination of many complex audio signals. With voiced speech, the speech signal is modified by either the oral cavity or the nasal cavity. These cavities act as resonators with pole and zero frequencies. Pole and zero frequencies are called formant and anti-formant frequencies ,respectively. These frequencies have their own amplitude and bandwidth. Voiced speech also produces a complex quasi-periodic pressure wave from an interruption in air flow caused by the vibration of the vocal folds. The frequency of impulses from the pressure wave is called the fundamental frequency. With purely unvoiced speech, since there is no vibration of the vocal folds,there is no fundamental frequency.Researchers believe that the most important signals generated by human speech are the formant frequencies and the fundamental frequencies. When synthesizing speech,

these signals greatly contribute to the naturalness of the speech. The formant frequency represents the shape of the sound that is formed by the vocal tract (oral cavity and the nasal cavity). Different sounds (vowels, nasals, etc.) within a language are distinguishable by their formant frequencies. The fundamental frequency determines the pitch of the voice. For example, women and children have a higher pitch (i.e. higher fundamental frequency)than men.Sounds created by humans are merely noise if the sounds do not have meaning. Sounds in speech production are categorized into units. These units can be as large as words or as minute as a phone. However, phonemes are the fundamental units of phonology. The definition of phonemes is the theoretical unit of sound that can distinguish words. The concatenation of phonemes produces the words in the language, i.e. changing a phoneme means changing the word. Phonemes are split into two major categories: vowels and consonants:-All vowels are voiced sounds while some consonants are voiced sounds and some are unvoiced sounds.With speech synthesis, the role of phonemes and diphones is to focus on sounds that the system should yield.

V .DESIGN AND IMPLEMENTATION OF TTS

A text-to-speech system 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*. 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 back-end—often referred to as the synthesizer – then converts the symbolic linguistic representation into sound, by a concatenative approach.The user writes the text either in Devnagri or Romanized script of Hindi language.The written text is processed only in Romanized script but also adaptable for Devnagri script. Such system for Hindi language is only possible when Devnagri is converted to Romanized script. Once the Romanized script is obtained, the task of text analysis and speech generation of TTS system is done.

In order to develop a TTS system, the bottom down approach deals with the designing of database. Preparing a database is the main theme of any TTS system. The database is designed purely on Hindi language but represented in Romanized script for the ease of platform compatibility. The database is formed by preparing various phonemes of Hindi language.A word document is interfaced by the user to write Devnagri script.

VI. SUMMARY AND CONCLUSIOS

• The area of speech generation had been the hardest speech technology area to obtain any viable degree of success. For more than 50 years researchers have struggled with the problem of trying to mimic the physical processes of speech generation.

• In spite of the best efforts of some outstanding speech researchers, the quality of synthetic speech generated by machine was unnatural most of the time and has been unacceptable for human use in most real world applications.

• TTS for Hindi language is developed for generation of speech for written text. The written text is either in Devnagri or Romanized script.

• By varying the values of these control variables within specific choices, an analysis of prosody with best possible case is determined. Starting from the least to most complex case of written text, the sentences are used for speech generation. The complexity rises from two character word to a sentence with all Hindi language *matra's.*

• With few limitations, the synthesis of speech for written text is well performed. The Speaker used for recording the sound files was untrained. And no sophisticated recording environment and instrument was used resulting some degradation in the prosody of synthesize speech signal.

• In many cases, the conversion of Devnagri to Romanized script leads to mismatching which also affects the prosody of synthesizes speech.

• There is a problem in preparing the database of some Devnagri phonemes which have same representation in Romanized script due to which they are represented by different letters.

• With such limitation, the synthesis of artificial speech is well performed along with the conversion of Devnagri to Romanized script.

• In the future, if speech recognition techniques reach adequate level, synthesized speech may also be used in language interpreters or several other communication systems, such as videophones, videoconferencing, or talking mobile phones.

ACKNOWLEDGEMENT

This thesis is the result of work carried out during the final semester of my course where by I have been accompanied and supported by many people. It is a pleasant aspect that I now have the opportunity to express my gratitude for all of them.

I express my deepest sense of reverence and indebtedness to the esteemed members of my Advisory Committee, Mrs. Reena Jain Asst. Professor . I am thankful to Dr. A.S Raghuvanshi, principal, Sobhasaria Engineering College Sikar, Rajasthan Technical University for providing necessary facilities to carry out the study.

REFERENCES

[1] Adell, J.; Escudero, D. and Bonafonte, A. 2012 'Production of filled pauses in concatenated speech synthesis based on the underlying fluent sentence', *Speech Communication, Barcelona.*

[2] Alias, F.; Formiga, L. and Llora, X. 2011 'Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept' *Speech Communication, Urbana, USA.*

[3] Bapat, V. A. and Nagalkar Lalit, K. 2008 'Phonetic Speech Analysis for Speech to Text Conversion', *IEEE region 10 coloqum & the 3rd International Conference on Industrial & Information system, Nagpur*.

[4] Chapell, D. and Hansen John, H.L. 2002 'A comparison of spectral smoothing methods for segment concatenation based speech synthesis', *Speech Communication, Durham, USA* .

[5] Chopra, D. 2011 'GAYATRI- A fast Hindi TTS system with input support for English Language', *International Journal of Information Tech & knowledge Management*, Mumbai.

[6] Cordoba, R. ; Montero, J.M. ; Gutierrez, J.M. ; Vallejo, J.A. ; Enriquez, E. and Pardo, J.M. 2002 'Selection of the most significant parameters for duration modeling in a Spanish text-to-speech system using neural networks' *Computer Speech and Language, France.*

[7] Dutoit, T. and Cernak, M. 2005 'TTSBOX: A MATLAB toolbox for teaching text-tospeech synthesis' *IEEE transaction system, Belgium.*

[8] Dhvani, 2001 TTS system for Indian Languages. *Project : http://dhvani.sourceforge.net.*

[9] Gillat A.M. 2008 'Numerical methods for engineers and scientists: an introduction with applications using MATLAB' *Publisher: Wiley; 1st edition.*

[10] Ki-Seung, L. and Richard, V. Cox 2002 'A segmental speech coder based on a concatenative TTS' *Speech Communication, USA.*

[11] Krstulovic, S.; Bimbot F. ;Olivier, B.; Delphine, C.; Dominique, F. and Odile, M.2006 'Optimizing the coverage of a speech database through a selection of representative speaker recordings' *Speech Communication, France*.

[12] Lawrence, R.R. and Juang, B.H. 1993 'Fundamental of Speech Recognition', *PTR Prentice Hall.*
Lazaridis, A. ; Mporas, I; Ganchev, T. ; Kokkinakis, G. and Fakotakis, N. 2011

[13] 'Improving phone duration modeling using support vector regression fusion' *Speech Communication,Greece*.