Processing of Devnagari Text to Speech Synthesis: A Review

Dr. Sunil Nimbhore[1], Dr. Suhas Mache[2]

[1]*UDOCS&IT, Dr.B.A.M.Univesity, Aurangabad (MH)-India.*
[2]*R. B. Attal College of Arts, Science and Commerce, Georai (MH)-India.*
nimbhoress@gmail.com, suhas.mache@yahoo.in

Abstract

This Paper describes a comprehensive survey related to the processing of Devanagari text for speech synthesis. In Human to Human Communication, Speech is primary and writing is a secondary method. It is desirable to have a similar mode of communication between Human-Computer interfaces. Therefore it is necessary to have speech interfaces to computers. TTS-synthesis is leading and more active research area. In this technology, natural language sentences in text form are converted into spoken form/production of speech sound (i.e. the form of the air stream). This paper describes the literature survey related to different types of synthesis system with their implementations and feature extraction techniques.

Keywords: Speech Synthesis, HMM-Synthesis, HNM-Synthesis, Linguistic Analysis, Prosody, Waveform Generation.

1. Introduction

Text-to-speech synthesis (TTS) is one of the computers assist interfaces which use natural language sentences or words are converted into the spoken form/waveform generation. The TTS system applications like in telecommunication services, aid to disabled people, education, talking toys, and multimedia, etc. The TTS-system basically into two module Natural Language Processing (NLP) module and a Digital Signal Processing (DSP) module as shown (Figure-1.) The given input text, the NLP module produces the phonetic transcription and its corresponding prosodic information. The DSP module converts information into the symbolic form to the speech signal. The DSP module laid to different approaches for speech synthesis [2]. Text-to-speech synthesis is an artificial production of human speech. The computer assists speech synthesis system is called a Speech Synthesizer. Synthesized speech can be created by concatenating pieces of recorded speech that are stored in the database is called as TTS-system [3].

The main objectives of speech synthesis on natural sounding totally depends on the prosodic model, the good prosodic model should be able to produce durational and intonation properties of natural speech.

The paper is organized as follows: In Section-1, the introduction part of speech synthesis is summarized. The TTS-System categorized into two components NLP-components and DSP-components. Section-2, it depicted the TTS front end (NLP-Component). The Section-3 describes the TTS-Back End (DSP-Component). Section-IV deals with some speech synthesis techniques like Formant Synthesis, Concatenative Synthesis, and HMM-Synthesis, Articulatory Synthesis, Unit Selection Synthesis, HNM-Synthesis. Section-V describes the speech database design. Section-VI summarizes application of speech synthesis used in the various area. The section-VII paper is concluded.

2. TTS Front End (NLP-Component)

In the Linguistic analysis, the following components are used to determine the specific text or sentences.

2.1. Text Pre-processing

It is responsible for influencing all knowledge regarding the text that is not purposely phonetic or prosodic. Text pre-processing is the end of sentence detection is also called Text Normalization. In the grammatical analysis, such as grammatical part-of-speech (POS) is assign for synthesis. Text normalization is the process of converting non-standard words (NSW) like numbers, abbreviations, titles, numerals etc., to natural pronounceable words. For speech synthesis can be produced by several methods are beneficial for synthesis.

2.1.1. Phonetic analysis

It transcribes lexical orthographic symbols into phonemic representations, possibly is a diacritic mark in order such as stress situation.

2.1.2. Prosodic analysis

It determines the proper intonation, verbal communication rate and amplitude for each phoneme in the transcription. [4][5]

2.1.3. Syntactic analysis

It is now necessary how the words fit into phrases and sentences, so to generate the sentence structure will be used to compute the prosody. The first step is called tagging involves to defining which syntactic category of each word belongs to (e.g. noun, verb, and adjective). The syntactic parser is used to build up a representation of each sentence structure [18]. Using this information a decision can be made about the stress is assigned to each part of each phrase.

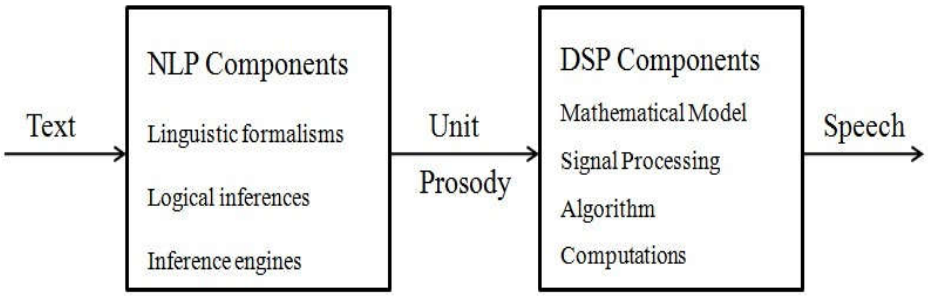

Figure 1. A Functional Block Diagram of TTS

3. TTS back-end (DSP -component)

The important qualities of the speech synthesis systems which are naturalness and Intelligible. The naturalness and intelligibility of the TTS system which is related to the output like human speech. In DSP components the statistical models are useful for speech synthesis techniques [6] as deliberated as follows:

3.1.1. Text Analysis

Text analysis converts the input text into a waveform/Speech Generation. Normalization of the text datasets is replaced by their corresponding whole words. This process naturally employs that large set of rules that try to take some Indian languages. The most exciting task in the text analysis portion is the phonological study. The syntactic and semantic analysis is used to understanding the content of the text.

3.1.2. Linguistic Analysis

Linguistic analysis is recognized as syntactic and semantic parsing. The linguistic analysis, which deals with Phrasing, Intonation, and duration, is the most important factor of speech, is called prosody of speech [21]. The linguistic tagging is used proper phonemes (verbs/noun/adjective). The following three terms are related to linguistic analysis;

Phrasing: The manually prepared or statistically trained phrase breaks the sentences.

Intonation: Intonation pattern, as they are called is different in different languages, the linker with stress and intonation. E.g.: 'She's very beautiful'.

Duration: It is related to the length of phonemes, air-pressure, voice which determines the frequency of vibration of the vocal cords. [22].

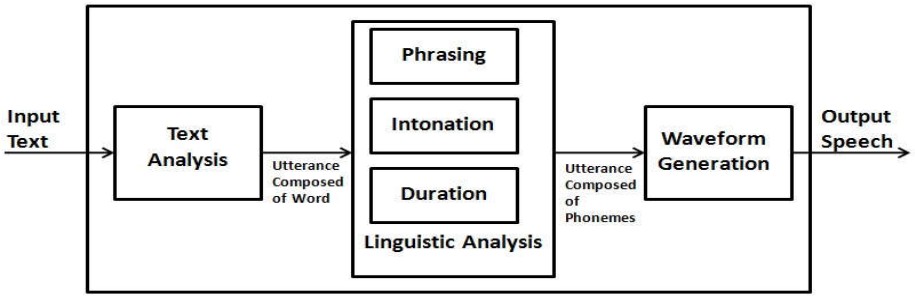

Figure 2. Block Diagram of Text-to-Speech Conversion

4. Speech Synthesis Techniques

The process of synthesizing speech is divided into two categories: (a) Analysis and (b) Synthesis. There are popular techniques of speech synthesis which are discussed in this section. The naturalness and Intelligible is an important quality of the speech synthesis system. The overall quality of the TTS-system is like human speech. The computer assists Marathi speech synthesis model is like natural human speech. The various speech synthesis techniques are described in the following sections

4.1.1. Formant Synthesis

Formant synthesis widely used in speech synthesis. It is based on the source and filter model of speech. The Marathi synthesized speech output is created using an acoustic model. The fundamental frequency (f0), expression, and noise levels parameters are varied over time to time so it's difficult to create a natural or artificial speech. The Rule-based synthesis is based on a set of rules used to determine the parameters to synthesize a preferred utterance using formant synthesizer.

Formant synthesis systems have advantages over Concatenative systems. The formant synthesizers which usually a small program than Concatenative system because they can't take a database of speech samples. This technique having control of prosodies, intonations, a variety of output speech [7].

4.1.2. Articulatory Synthesis

The articulatory synthesis is the first technique approaches to speech synthesis. In this approach, machines are made to mimic humans in the same way as humans to produce speech, i.e., the human speech production system. In, the vocal and nasal tracts are treated as tubes that are attached with closures for articulators such as the tongue, jaw, and lips, in Human Production Speech System.

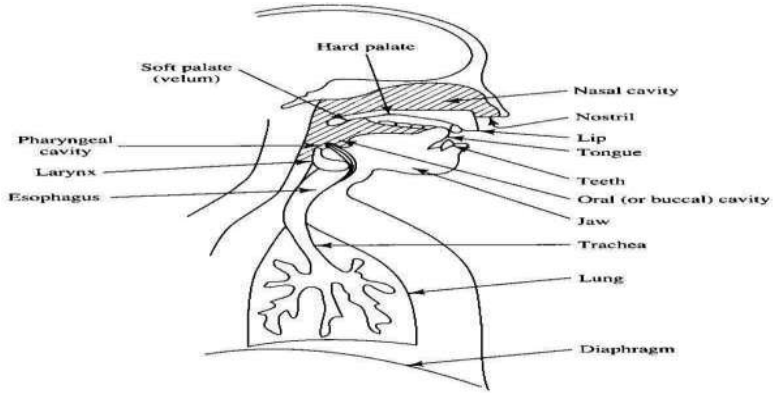

Figure 3. View of human speech production system [8].

In this model that used hand control of a vocal track made of leather, with bellows for lungs, auxiliary holes for nostrils, and reeds and other mechanisms for vocal cords vibration and turbulence creation. These mechanical analogs could produce vowels, nasals, and a few simple utterances. [9]

4.1.3. Concatenative Synthesis

Concatenative synthesis is the dictionary-based approach. Concatenative synthesis is based on the concatenate of segments of pre-recorded natural human speech or (stringing together). Concatenative synthesis produces intelligible & natural sounding synthetic speech, usually close to the voice characteristics of a real person. However, Concatenative synthesizers are limited to one speaker & one voice [10]. A difference between natural variations in speech signals and the nature of the automated techniques for generating the waveform of speech [11]. There are main three sub-types of Concatenative synthesis: unit selection synthesis, Diphone synthesis and domain-specific synthesis [12].

4.1.3.1. Unit Selection Synthesis

The Unit selection synthesis is known as corpus-based concatenative synthesis. It solves the problem by storing in the speech of unit records multiple instances of each unit with varying prosodies. The speech unit that is matched with a close target of the prosody. It concatenated with that prosodic modification is needed on the selected speech unit. In case of automatic unit selection synthesis. [20]

4.1.3.2. Single Diphone-Based Synthesis

Diphone is contained in the speech database. The Diphone synthesis speech database containing all the Diphone (sound-to-sound) occurring in the text. A Diphone is a speech segment that starts in the middle of the stable part of a phoneme and ends in the middle part of the next phoneme. These single example Diphones are used in synthesis, with the desired prosody for articular context realized via digital signal processing. Prosody modeling plays a vital role in this technique. Diphone uses wide-ranging digital signal processing, which reduces the distortion from concatenation points of the synthesized speech signal. [13]

4.1.3.3. Automatic Unit Selection Synthesis

This synthesis is a speech database is designed such that multiple instances of each unit are available in various prosodic and phonetic contexts. The naturalness of these units is what contributes to the naturalness of such synthesizers. [14]
The unit selection synthesis most commonly phonemes are used for synthesis because they are the normal phonology of speech. Some linguistic modules or analysis which is described in section-4.

HNM-Synthesis

In HNM based synthesis model is a parametric model, it's very easy to modify the prosodic features like intonation, stress or rhythm within a good quality [15]. In HNM model as composed of harmonic and noise part. The harmonic part is representing the prosodic component of speech and Noise part represent non-prosodic components. These two parts separated in the frequency domain called as voiced frequency. The Harmonic plus Noise Model (HNM) into the Hidden Markov Model Speech Synthesis System. [16][17]

A. HMM- Synthesis

A Hidden Markov Model (HMM) is simply a Markov in which the states are hidden. HMM-based synthesis also called Statistical Parametric Synthesis. The HMM speech system model having parameters like frequency spectrum, fundamental frequency, and duration resp. The Speech waveform generated from HMMs based on the maximum probability measure. Hidden Markov Model is also used to recognition word for speech in the text to speech system [18]. The use HMMs to generate "parts of speech" (POS) this tagging is lead disambiguating homographs.

This technique is quite successful in many cases. For example let $P(A|W)$ be the probability of given word sequence W, generating acoustic evidence A. Recognition algorithm must identify the word sequence that maximizes P (W|A) by Bay's formula:

$$P(W|A) = [P(W)P(A|W)]/P(A) \quad (1)$$

Where P (W) is a priori probability that the word W is uttered. Average probability is acoustic evidence A is observed and $P(A|W)$ is the probability that when W is uttered. So, our problem is to maximize $P(W)P(A|W)$. $P(A|W)$ Corresponds to the acoustic level recognition and $P(W)$, the language model, using probability theory.

$$P(W) = \prod nP(W_i|W_1, W_2, \ldots..W_{i-1}) \quad (2)$$

Here, P(Wi|W1, W2…..Wi-1) is the probability that Wi will be spoken, given that words W1, W2…..Wi-1 was said before and n is the number of a word taken in sequence. An ESTIMATE MODIFIED calculations are to take the sequences with the same two words at the end an 'equivalent' and calculate probability: [19]. $P(W) = \prod nP(W_{i|Wi-2}, W_{i-1})$ (3)

5. Database Design

The database acquisition for the Concatenative system is used predefined units of Marathi vowels, consonants, numbers, word and sentences which are written in Devnagari script. The Marathi Speech database of these units which are design for the purpose of this system.

6. Application of Speech Synthesis

The applications of speech synthesis systems are inclusive extending for HCI, telecommunications, talking books, language education and aid to handicapped persons.

A. News Reader

The newsreader extracts the new items from major news portals, and does a lot of text processing to clean up the text, before giving it to the speech generation component.

B. Screen Reader for Visually Impaired

This screen reader application is helpful for visually handicapped or visually impaired persons. It aids the person to navigate through the screen and reads out the news items and articles available in desired language scripts. The TTS-system is very helpful for impaired people because they feel like taking helped by a teacher.

C. Telecommunication

The synthesized speech is used for all kind of telephone inquiry systems. Synthetic speech is the key factor in voice mail systems.

D. Multimedia

Nowadays multimedia is new applications of speech synthesis to create animated stories, poems in various languages. Text-to-Speech system tool is like a Computer Aided Learning system can provide to learn a new language.

E. Education

In the interactive voice-based speech synthesis models are used in the educational field. The computer-based speech synthesizer is to teach the spelling and pronunciation of various languages.

7. Conclusion

This paper has been signified as a survey of speech synthesis. It is related to synthesis techniques which are used for Devanagari text processing. The literature survey of text to speech synthesizers are implemented in different languages which are listed in Table-I. This survey of speech synthesis techniques is a very challenging task for implementing a different speech synthesis system for different languages. This literature survey of research papers presents the theoretical and technological speech synthesizer.

Nowadays most powerful technique is the concatenation method because it is simple and easy to implement based on the available tools and environments. Which have emerged another one technique is unit selection synthesis it means the number of utterances for the analysis in units. Such synthesis techniques are permissible for more natural-sounding for generation of speech. The speech synthesis it uses Hidden Markov Model as a statistical method that allows more variations on the recorded data set. The Hidden Markov model-based speech synthesis systems become leading on the TTS. It may be needful the smaller development. This literature survey of TTS-System which are applying the various feature extraction techniques in throughout the research.

Acknowledgment

Authors are thankful for the head, Department of Computer Science & IT, Dr. B. A. M. University, Aurangabad for providing the infrastructure and lab facilities.

Table 1. Literature Review of TTS-System with their Feature Extraction Techniques [23, 24, 25, 28]

References

Sr. No.	TTS- System	Authors	Techniques	Implementation for TTS
1.	Leather Resonator	Von Kempelen, et. al.	Mechanical Synthesizer	Vocal tract Produces (Vowels, Glides, Nasals)
2.	Acoustic- Mechanical Speech Machine	Wolfgang von Kempelen	Mechanical Synthesizer	Produces Single Sound and Sound Combinations.
3.	Speech Organs	Charles Wheatstone	Mechanical Synthesizer	It sings the God Save the Queen
4.	Electrical Synthesis Device	Stewart	Electrical Synthesizer	It is used for the two formants.
5.	Natural Vocal Tract	Riesz et. al.	Electrical Synthesizer	Vocal tract Produces the Natural Vowels.
6.	VODER	Homer Dudley et.al.	Electrical Synthesizer	Human controlled through complex Keyboard & Pedal
7.	Pattern Playback	Franklin Cooper et. al.	Electrical Synthesizer	It's used for investigating speech spectrograph
8.	PAT, "Parametric Artificial Talker"	Walter Lawrence's et. al.	Formant Synthesizer	It consists of i/p signal was noise or buzz.
9.	Copying a natural sentence using PAT	Walter Lawrence's et. al.	Formant synthesizer	Phonemics Synthesis by rule
10.	First full text-to-speech system.	Noriko Umeda et. al.,	Full Automatic System	It produced the text to the speech signal.
11.	Comparison of synthesis	John Holmes et. al.	Formant synthesizer	Natural Sentences are Compared
12.	Articulatory Synthesis	James Flanagan and Kenzo Ishizaka	Articulatory Synthesis	It Produced sentences
13.	MIT MI Talk,	Allen, Hunnicut, Klatt et.al.	Formant Synthesis	Filter to create each formant.
14.	Linear-prediction analysis (LPC) and resynthesis	Richard Wiggins et.al.	Re-Synthesis of Speech	Introducing the Instruments Speak 'n' Spell toy.
15.	DEC Talk commercial system	Dennis Klatt et. al.	Articulatory Synthesis	Formed by Basic Digital Equipment
16.	Pitch Synchronous Overlap Add. (PSOLA)	Moulines and Charpentier et.al.	Signal Synthesis	It is Time Scale and Pitch Scale Modifications
17.	Proverb and HIDIFIX	Davnovan et. al.	Signal Synthesis	Pitch and Duration based Commercial System
18.	TD-PASOLA	Kortekaas et.al.	Time Domain Version	It's Used for Computational efficiency
19.	Text-To-Speech System	AT & T Lab	Concatenative Synthesis	It Produced Prosody
20.	MIT Talk-System	By Dennis Klatt at Bell Lab.	NLP Techniques	First multilingual language-independent systems.
21.	Festival TTS system	By Alan Black and Paul Taylor	LPC coefficient Techniques	This system is developed for different aspects.
22.	Infovox	Ljungqvist et al.	Formant Synthesis	Multilingual text-to-speech products it's a commercial version.
23.	Cyber Talk	Panasonic Technologies, Inc. (PTI), USA	Rule-based formant synthesis	Text-to-speech synthesis system for English

[1] T. Dutoit "An Introduction to text-to-speech synthesis", Kluwer Academic Publishers, 1997.
[2] Aimilios Chalamandaris, Sotiris Karabetsos, Pirros Tsiakoulis, Member and Spyros Raptis, "A Unit Selection Text-to-Speech Synthesis System Optimized for Use with Screen Readers", IEEE, Vol. 56, No. 3, August 2010.
[3] Hala ElAarag, Laura Schindler, "A Speech Recognition and Synthesis Tool", ACMSE 2006.
[4] K.Kiran Kumar, K.Sreenivasa Rao and B.Yegnanarayana "Duration Knowledge for Text–to–Speech Conversion System for Telugu", [ICKBCS].
[5] G. L. Jayavardhana Rama, A. G. Ramakrishna, M Vijay Venkatesh, R. Murali Shankar, "Thirukkural–A Text to Speech Synthesis System", 2000.
[6] P. Prathibha, A. G. Ramakrishnan, R. Muralishankar, Thirukkural II - A Text-to-Speech Synthesis System", 2002.
[7] Bulyko, Ivan and Ostendorf, Mari, "The Impact of speech recognition on speech synthesis", Proceedings of IEEE workshop on Speech synthesis, 2002.
[8] B. Kroger, "Minimal Rules for articulatory Speech Synthesis," Proceedings of EUSIPCO92, pp.331-334, 1992.
[9] Venugopalkrishnan Y. R, Methods for improving the Quality of Syllable-Based Speech Synthesis for Indian Languages, Ph.D. thesis, 2009.
[10] Mr.S.D.Shirbahadurkar, Dr.D.S.Bormane, "Speech Synthesizer Using Concatenative Synthesis Strategy for the Marathi language (Spoken in Maharashtra, India)", ACEEE, 2009.
[11] Alex Acero, "An Overview of Text-to-Speech Synthesis', Speech Technologies Group Microsoft Corp. IEEE, 2000.
[12] Lawrence Rabiner, "Speech Recognition in Machine: A Review" International journal of computer science and Information Security 2009.
[13] E.Veera Raghavendray, Kishore Prahallad, "A Multiline dual Screen Reader in Indian Languages", IEEE, 2010.
[14] Masatsune Tamura, Norbert Braunschweiler, Takehiko Kagoshima and Masami Akamine, "Unit Selection Speech Synthesis Using Multiple Speech Units at Non-Adjacent Segments for Prosody and Wave Form Generation", ICASSP, IEEE, 2010.
[15] Y. Stylianou, "Modeling Speech Based on Harmonic Plus Noise Models," Springer, 2005.
[16] C. Hemptinne. "Integration of the Harmonic plus Noise Model into the Hidden Markov Model-Based Speech Synthesis System (HTS)," Master Thesis: IDIAPLauzane, Suisse 2006.
[17] Y. Stylianou, "Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification, "Ph.D. thesis, Ecole Nationale upérieure des Telecommunications, Paris, January 1996.
[18] Anannd Arokia Raj, Tanuja Sarkar, Satish Chandra Pammi, et. al., "Text Processing for Text-to-Speech System in Indian Language", ISCA Workshop on Speech Synthesis, August 2007.
[19] Alan W Black, Heiga Zen, Keiichi Tokuda, "Statistical Parametric Speech Synthesis", 2006.
[20] S.P.Kishore, Alan W Black, Rohit Kumar, and Rajeev Sangal, "Experiments with Unit Selection Speech Database for Indian Languages", I2IT, Hyderabad.
[21] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, M. Plump, "Recent improvements on Microsoft's Trainable Text-to-Speech System-Whistler", Microsoft Research, 1999.
[22] Youcef Tabet and Mohamed Boughazi, "Speech Synthesis Techniques: A Survey. Pages 67-70, (WOSSPA) IEEE, 2011.
[23] Dan Jurafsky, "Lecture Notes on Richard Sprout's Slides Speech Recognition, Synthesis, and Dialogue".
[24] Igor Sz oke, FIT BUT Brno, "Lecture Notes on Speech Synthesis".
[25] T. Mizutani and T. Kagoshima, "Concatenative speech synthesis based on the plural unit selection and fusion method (in English)," IEICE Trans., vol. E88-D, no. 11, pp. 2565–2572, 2005.
[26] A. J. Hunt and A. W. Black, "Unit selection in a Concatenative speech synthesis system using a large speech database," Proc ICASSP-96, pp. 373–376, 1996.
[27] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigen voices for HMM-based speech synthesis," in Proc. ICSLP, 2002, pp. 1269–1272.
[28] G. Gidda Reddy, "Speech Analysis-Synthesis for Speaker Character Modification", IIT Bombay, Nov-2004.
[29] Online, "HMM-based Speech Synthesis System (HTS)", http://hts.sp.nitech.ac.jp/ visited by 2011.
[30] Online, "Text-to-Speech bibliography", http://texttospeech.org/ visited by 2011.
[31] R.K. Bansal, and J.B. Harrison, "Spoken English for India, A MANUAL OF SPEECH AND PHONETICS", Orient Longman, 1972.
[32] Rabiner, Lawrence and Schafer, Ronald, Digital Processing of Speech Signals, Prentice-Hall, 1978.
[33] John Proakis and Dimities Manolakis, Digital Signal Processing 3rd Edition, 1996.
[34] Lawrence R. Rabiner and Ronald W. Schafer," Introduction to Digital Speech Processing", Now Publishers, USA, 2007.