Different Approaches of Speech Syntheses and Text to Speech Synthesis Systems in Indian Languages

Archana N A
Department of Computer Science & Engineering
Vidya Academy of Science and Technology
Thrissur, India

Vidya M
Department of Computer Science & Engineering
Vidya Academy of Science and Technology
Thrissur, India

Abstract

Text to speech system, also known as speech synthesis is the process of generating spoken language by machine on the basis of written input. The purpose of this paper is to introduce the area of speech synthesis by giving an overview of the three main methods of the speech synthesis as concatenative synthesis, articulatory synthesis and formant synthesis. To build a natural speech synthesis system, it is essential that the text processing component produce a sequence of phonemic units corresponding to the arbitrary input text. This paper discusses Font to Akshara mapping, pronunciation rules for aksharas and text normalization in the context of building text to speech systems in Indian languages. The paper is organized as follows: Section 1 provide an introduction. Section 2 gives different types of speech synthesis Section 3 provides an overview of the text to speech system in Indian languages and Section 5 concludes the manuscript.

Keywords: text to speech, concatenative synthesis, phonemes, Natural Language Processing

I. INTRODUCTION

The automatic conversion of written text into speech would prove a vitally useful technology with many commercial and humanitarian applications. Speech is one of the most natural ways for humans to communicate with one another; enabling a computer to convey information in this manner has marked advantages over written text.

Speech synthesis would be valuable as an assistive technology but it can also be applied to more general commercial applications. Amongst the typical uses of the technology are:

A. Talking aids for the Vocally Handicapped
Any person who is unable to speak but can use a typewriter or similar interface has the potential to use a text-to-speech system to provide themselves with a voice. The synthesised voice can be tailored to the person with specific characteristics to retain individuality.

B. Reading aids for the Blind
The visually impaired can benefit tremendously from text-to-speech technology. Text-to-speech software would enable input text to be generated to spoken words, and allow access to information available online.

C. Training and Educational Aids
With regard to many cognitive activities, it is known that speech has several advantages over written language. Virtual teachers, contributing to a distance learning course, for example, could teach on-line tutorials. This can be particularly advantageous in situations where the presence of a real teacher can be embarrassing for the student, as has been noted for sufferers of dyslexia.

D. Remote Access to Online Information
Any written information that is stored online, for example electronic mail, news items, directory enquiries, can all be accessed aurally by means of a speech synthesizer.

The objective of a text to speech system is to convert an arbitrary given text into a corresponding spoken waveform. Text processing and speech generation are two main components of a text to speech system. The objective of the text processing component is to process the given input text and produce appropriate sequence of phonemic units. These phonemic units are realized by the speech generation component either by synthesis from parameters or by selection of a unit from a large speech corpus. For natural sounding speech synthesis, it is essential that the text processing component produce an appropriate sequence of phonemic units corresponding to an arbitrary input text.

Firstly, this paper illustrates the three main methods of speech synthesis. The second section provides information about text processing for text-to-speech systems in Indian languages.
II. SPEECH SYNTHESIS

In latter part of the 18th century, a Hungarian nobleman, Wolfgang von Kempelen, developed a speaking machine based upon his observations of human speech production. The machine consisted of a pressure chamber in place of the lungs, a vibrating metal reed representing the vocal folds, and a leather sack for the vocal tract, which could be manipulated to produce different vowel sounds. The inclusion of various different models of the tongue and lips made it possible to produce plosives. It was reported that the talking machine could speak whole phrases in French and Italian.

Before the advent of electronic systems and computers, it was only through mechanical means like Kempelen’s that speech-like sounds could be synthesised. Since the first third of the 20th century, the deeper understanding of the human speech production system coupled with technological advancements allowed for the development of electronically-based speech synthesis. The following subsections describe the main principles of the three most commonly used speech synthesis methods: formant synthesis, concatenative synthesis, and articulatory synthesis.

A. Formant Synthesis

This is the oldest method for speech synthesis, and it dominated the synthesis implementations for a long time. Nowadays the concatenative synthesis is also a very typical approach. Formant synthesis is based on the well-known source-filter model which means that the idea is to generate periodic and non-periodic source signals and to feed them through a resonator circuit – or a filter – that models the vocal tract. The principles are thus very simple, which makes formant synthesis flexible and relatively easy to implement. In contrast to the methods described below, formant synthesis can be used to produce any sounds. On the other hand, the simplifications made in the modeling of the source signal and vocal tract inevitably lead to somewhat unnatural sounding result.

Spectrograms taken of speech signal reveal the dynamic nature of the frequency content of human speech. As speech is produced, the movements of the jaw, tongue and other articulators enhance certain resonances that change according to the vocal tract geometry. These resonant regions exhibited by the vocal tract are called formants. Formant synthesis seeks to mimic human speech by artificially creating the movements of these frequency resonances.

![Fig. 1: Sample formant structures for the vowel /IY/, /AA/ and /UW/.](image)

The parameters necessary for driving formant synthesis models can be derived from mathematical analyses of continuous speech, for example by Fourier analysis or by linear prediction methods.

Though the theory upon which it is based is a simplification of the actual human speech production process, by making functional approximations to these phenomena, formant synthesisers can produce very high quality speech. Indeed, in the early 70s, Holmes showed that his formant synthesiser could be used to reproduce a nearly perfect duplicate of the male voice, but the determination of the complex control tables took months of manual adjustment.

B. Articulatory Synthesis

While formant synthesis attempts to create the spectral features of a particular speech sound directly, articulatory synthesis does so by modelling the geometry of the vocal tract that would re-create a specific spectrum. By approaching speech production from this physical modelling standpoint, it has a direct relation to the human vocal system. Thus, it has been conjectured that this method will eventually lead to a complete vocal structure model, capable of reproducing all the sounds utter-able by the human speech reproduction system.

Articulatory synthesis is by far the most complicated in regard to the model structure and computational burden. The idea in articulatory synthesis is to model the human speech production mechanisms as perfectly as possible. The implementation of such a system is very difficult and therefore it is not widely in use yet. Experiments with articulatory synthesis systems have not been as successful as with other synthesis systems but in theory it has the best potential for high-quality synthetic speech. For example, it is impossible to use articulatory synthesis for producing sounds that humans cannot produce (due to human physiology). In other synthesis methods it is possible to produce such sounds, and the problem is that these sounds are usually perceived as undesired side effects. The articulatory model also enables more accurate transient sounds than other synthesis techniques.

Articulatory synthesis systems contain physical models of both the human vocal tract and the physiology of the vocal cords. It is common to use a set of area functions to model the variation of the cross-sectional area of the vocal tract between the larynx and the lips. The principle is thus similar to the one that has been seen within the acoustic tube model. The articulatory model involves a large number of control parameters that are used for the very detailed adjustment of the position of lips and tongue, the lung pressure, the tension of vocal cords, and so on.

Articulatory synthesis completes a full circle with the type of approach taken by von Kempelen hundreds of years ago. However, instead of physically manipulating a mechanical model of the human speech production system to produce utterances, speech is calculated as the output of a virtual model simulated on a computer. Typically, the vocal tract is divided into many small sections...
whose dimensions collectively determine the resonant characteristics of the vocal tract (Fig. 2). Mimicry of speech is achieved by dynamically changing the virtual shape and sizes of these segments according to the corresponding articulatory movements.

Vocal tract dimensions have been obtained by measurements of x-rays and other special laboratory methods, but ideally these parameters would be derived from natural speech, in a manner similar to the parameter determination of formant synthesisers. This, however, is a non-trivial problem, afflicted by problems of articulatory ambiguity. As such, a method to obtain articulatory parameters from a speech signal remains an on-going research direction.

**C. Concatenative Synthesis**

This is the so-called cut and paste synthesis in which short segments of speech are selected from a pre-recorded database and joined one after another to produce the desired utterances. In theory, the use of real speech as the basis of synthetic speech brings about the potential for very high quality, but in practice there are serious limitations, mainly due to the memory capacity required by such a system. The longer the selected units are, the fewer problematic concatenation points will occur in the synthetic speech, but at the same time the memory requirements increase. Another limitation in concatenative synthesis is the strong dependency of the output speech on the chosen database. For example, the personality or the affective tone of the speech is hardly controllable. Despite the somewhat featureless nature, concatenative synthesis is well suited for certain limited applications.

It is possible to reproduce speech messages by playing back recordings of spoken words in the correct order to generate the desired message. Such data-driven systems are called concatenative synthesisers, as they concatenate speech segments to produce utterances. This was the type of method employed by the UK telephone network’s speaking clock, introduced in 1936. The various phrases and words were carefully recorded with regards to pitch, stress and other prosodic elements to ensure that reasonable fluency and naturalness are retained at synthesis time. Similar schemes have been employed by transportation networks, meteorological services and other such systems requiring a limited set of words.

The concatenation of whole words can produce extremely natural sounding results but they can become restricted by their lack of flexibility. They are limited by the memory available for storage but also by the problems inherent in recording and editing the new words. However, it can be possible to overcome these difficulties by concatenating speech segments which are less than a word in length.

The size of sub-word unit to use for purposes of concatenation needs to be chosen with care. Experiments using individual phones as the building blocks of speech demonstrated how normal speech sounds are heavily influenced by its neighbouring utterances. The coarticulation phenomenon results from the movement of the articulators approximating target positions rather than exactly reaching them. Recordings played back with any regard for this co-articulation have been judged to be extremely difficult to listen to, as a result of the wrong intonation and rhythm.

As speech sounds generally consist of a steady-state region as well as transitional sections, researchers have been able to overcome the problem of co-articulation somewhat by concatenating phonemic transitions. However, there still remains a high degree of variation at the steady-state regions in these segments, according to the identity of the adjacent phones.

Modern concatenative systems utilize extremely large speech corpuses from which to draw their speech segments. Such systems are called unit-selection concatenative synthesisers, emphasising perhaps that the key to synthesis isn’t the concatenating of speech segments but the selection of segments with the minimum of joins necessary, (Fig. 3). Specially designed cost functions determine which segments are chosen. By reducing the amount of digital manipulation of the signal, a high degree of naturalness can be achieved.
More recently, it has been suggested that as concatenative systems essentially ignore any information about the speech production mechanism, improved systems can be devised if the method of concatenative synthesis enters the articulatory domain. By eliminating articulatory discontinuities, it is hoped the imperfect joins that persist due to coarticulation can be rectified and more natural results be achieved.

III. TEXT TO SPEECH SYSTEMS IN INDIAN LANGUAGES

The area of speech processing has gone to the extent of synthesizing natural speech that highly resembles a human voice. The idea behind is so simple. The system is fed with a human voice, processed and stored as smaller units. When the target text is given for converting into speech, these smaller units are concatenated to build a continuous speech. This technique is called as Concatenative Speech synthesis. The processing of the human voice into smaller units is called the Speech Corpus development. The corpus can either be general purpose or application specific, based on the type of Speech Processing system. To build a corpus, a text file is first selected, which has phonetically and prosodically rich sentences. The text file may be a News bulletin, forum interviews, every-day conversations in an organization, conversation in road traffic, etc. Then, with the help of a native speaker, who may be a news reader, this text file is made to be read and recorded. This recording is the corresponding speech file

A speech file is an audio file whose size generally ranges from several minutes to hours. This speech file is the phonetic representation of the selected text file. The speech file is processed in different levels viz., paragraphs, sentences, phrases, words, syllables and phones, which are called the speech units of the file. Many researches have been carried out using these units as the basic units of processing. Research in the area of speech synthesis have an importance in many applications, including information retrieval services over telephone such as banking services, public announcements at places like train stations and reading out manuscripts for collation, reading emails, faxes and web pages over telephone and voice output in automatic translation systems.

A. Nature of Indian Language Scripts

The languages of India have a common phonetic base. One does not use the term "alphabet" to refer to the set of letters with which the script is written. Instead, the set is called "Aksharas". Very Simply, an akshara refers to a sound. Sounds heard in spoken words are built up from the basic set of sounds represented by the vowels and consonants of the language. The scripts in Indian languages have originated from the ancient Brahmi script. The basic units of the writing system are referred to as Aksharas. The properties of Aksharas are as follows: (1) An Akshara is an orthographic representation of a speech sound in an Indian language; (2) Aksharas are syllabic in nature; (3) The typical forms of Akshara are V, CV, CCV and CCCV, thus have a generalized form of C*V.

In all Indian languages, an akshara is pronounced the same way regardless of its position within a word, unlike in English where the pronunciation varies widely, depending not only on the word but also on the location of the letter within the word. The shape of an Akshara depends on its composition of consonants and the vowel, and sequence of the consonants. In defining the shape of an Akshara, one of the consonant symbols acts as pivotal symbol (referred to as semi-full form). Depending on the context, an Akshara can have a complex shape with other consonant and vowel symbols being placed on top, below, before, after or sometimes surrounding the pivotal symbol (referred to as half-form). Thus to render an Akshara, a set of semi-full or half-forms have to be rendered, which in turn are rendered using a set of basic shapes referred to as glyphs. Often a semi-full or half-form is rendered using two or more glyphs, thus there is no one-to-one correspondence between glyphs of a font and semifull or half-forms.

B. Convergence and Divergence

There are 23 official languages of India, and all of them except English and Urdu share a common phonetic base, i.e., they share a common set of speech sounds. While all of these languages share a common phonetic base, some of the languages such as Hindi, Marathi and Nepali also share a common script known as Devanagari. But languages such as Telugu, Kannada and Tamil have their own scripts. The property that makes these languages separate can be attributed to the phonotactics in each of these languages rather than the scripts and speech sounds. phonotactics is the permissible combinations of phones that can co-occur in a language.

C. Digital Storage of Indian Language Script

Another aspect of diversion of electronic content of Indian languages is their format of digital storage. Storage formats like ASCII (American Standard Code for Information Interchange) based fonts, ISCII (Indian Standard code for Information Interchange) and Unicode are often used to store the digital text data in Indian languages. The text is rendered using some fonts of these formats. There is a chaos as far as the text in Indian languages in electronic form is concerned. Neither can one exchange the notes in Indian languages as conveniently as in English language, nor can one perform search easily on texts in Indian languages available over the web. This is because the texts are being stored in ASCII font dependent glyph codes as opposed to Unicode.

The glyph coding schemes are typically different for different languages and within a language there could exists several font-types with their own glyph codes (as many as major news portals in a language). To view the websites hosting the content in a particular font-type, these fonts have to be installed on local machine. As this was the technology existed before the era of Unicode and hence a lot of electronic data in Indian languages were made and available in that form.

D. Phonetic Transliteration Scheme for Digital storage of Indian Language Scripts

To handle diversified storage formats of scripts of Indian languages such as ASCII based fonts, ISCII (Indian Standard code for Information Interchange) and Unicode etc, it is useful and becomes necessary to use a meta-storage format. A transliteration
scheme maps the Aksharas of Indian languages onto English alphabets and it could serve as metastorage format for text-data. Since Aksharas in Indian languages are orthographic represent of speech sound, and they have a common phonetic base, it is suggested to have a phonetic transliteration scheme such as IT3. Thus when the font-data is converted into IT3, it essentially turns the whole effort into font-to-Akshara conversion.

E. Identification of Font-Type
The widespread and increasing availability of textual data in electronic form in various font encoded form in Indian languages increases the importance of using automatic methods to analyze the content of the textual documents. The identification and classification of the text or text documents based on their content to a specific encoding type (specially font) are becoming imperative. Previous works were done to identify the language and later to identify the encodings also. Most of them N-gram based modeling technique. It may be helpful to make the difference clear here, the term refers a ‘glyph’ and the document refers the ‘font-data (words and sentences) in a specific font-type’. The Term Frequency - Inverse Document Frequency (TF-IDF) approach is used to weigh each term in the document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among glyph-sequence, font-data and font type. It may be helpful to make the difference clear here, the term refers a ‘glyph’ and the document refers the ‘font-data (words and sentences) in a specific font-type’. Here the glyph-sequence means unigram (single glyph), bigram (“current and next” glyph) and trigram (“previous, current and next” glyph) etc.

F. Conversion of Font-Data
By font conversion we mean here the conversion of glyph to grapheme (akshara). So we want to make clear about glyph and grapheme. A character or grapheme is a unit of text, whereas a glyph is a graphical unit. In graphonomics, the term glyph is used for a non-character, i.e: either a sub-character or multi-character pattern.

Font-data conversion can be defined as converting the font encoded data into required phonetic transliteration scheme like IT3 based data. A generic framework has been designed for building font converters for Indian languages based on this idea using glyph assimilation rules. The font conversion process has two phases, in the first phase we are building the Glyph-Map table for each font-type and in the second phase defining and modifying the glyph assimilation rules for a specific language.

G. Building Pronunciation Models for Aksharas
Having converted the font-data into Aksharas, the next step is to obtain appropriate pronunciation for each of the Aksharas. As noted earlier, Aksharas are orthographic representation of speech sounds and it is commonly believed or quoted that there is direct correspondence between what is written and what is spoken in Indian languages, however, there is no one-to-one correspondence between what is written and what is spoken. Often some of the sounds are deleted such as Schwa deletion in Hindi. Schwa is the default short vowel /a/which is associated with a consonant, and often it is deleted to aid in faster pronunciation of a word. Similarly there exists exceptions for Bengali and Tamil. There are attempts to model these exceptions in the form of the rules, however, they are often met with limited success or they use linguistic resources such as Morph analyzer. Such linguistic resources may not always be available for minority languages. Thus we had built a framework based on machine learning techniques where pronunciation of Aksharas could be modeled using machine learning techniques and using a small set of supervised training data.

H. Normalizing of Non-Standard Words
Unrestricted texts include Standard Words (common words and Proper Names) and Non-Standard Words (NSWs). Standard Words have a specific pronunciation that can be phonetically described either in a lexicon, using a disambiguation processing to some extent, or by letter-to-sound rules. In the context of TTS the problem is to decide how an automatic system should pronounce a token; even before the pronunciation of a token, it is important to identify the NSW-Category of a token.

IV. CONCLUSION
In this paper, the three main approaches to speech synthesis are described. While formant and articulatory synthesizes offer perhaps more flexibility than concatenative synthesizers, data-based methods are preferred because of their level of naturalness, seemingly unattainable by the rule-based versions. In this paper we also discussed about Text to speech synthesizer. It is observed that the development of TTS in Indian languages is a difficult task.

REFERENCES