Text To Speech

A Project Report submitted in partial fulfilment of the requirement for the award of the degree of
Bachelor of technology
In
Computer Science & Engineering

Under the Guidance of
Prof. Dr. Debaprasad Mukherjee

Submitted By:
SIDDHANT SENGAR, BALENDU KUMAR, SHANTANU SARKAR, NITIN KUMAR MADHESHIYA

Department of Computer Science & Engineering
Dr. B.C. Roy Engineering College
Fuljhore, Jemua Road, Durgapur - 713206
West Bengal, India

ACKNOWLEDGEMENT

Many people have helped to create this project and each of their contribution has been valuable. The timely completion of this project on Text To Speech is mainly due to the interest among us and of Prof. Dr. Debaprasad Mukherjee who has not only been a guide but also a good motivator. We would also like to extend our appreciation to Department of Computer Science & Engineering for giving us the chance to make a project on the above defined topic and providing us an ever-respected teacher who has contributed greatly by going through the whole document and giving valuable suggestions for its improvement. Our special thanks to our parents for their encouragement and unstinted support throughout the project.

We wish to thank our friends for being reviewers and criticizing which helped us in further improvement of this project.

(Team Members Name) Date: 05/05/2016

1. SIDDHANT SENGAR
2. BALENDU KUMAR
3. SHANTANU SARKAR
4. NITIN KUMAR MADHESHIYA

CERTIFICATE OF APPROVAL

This is to certify that, Balendu Kumar, Nitin Kumar Madheshiya, Shantanu Sarkar, Siddhant Sengar, students in the department of Computer Science and Engineering, worked on the project entitled Text To Speech**.**

I hereby recommend that the report prepared by them may be accepted in partial fulfillment of the requirement of the Degree of Bachelors of Technology in the Department of Computer Science and Engineering, Dr. B. C. Roy Engineering College, Durgapur.

………………………………………….
Prof. Dr. Debaprasad Mukherjee
Project Mentor
Department of Computer Science and Engineering
Dr. B. C. Roy Engineering College, Durgapur

.   .

Forwarded by:

………………………………
Head, Dept. of Comp. Sc. & Engg,
Dr. B. C. Roy Engineering College, Durgapur

CERTIFICATE OF APPROVAL

This is to certify that Balendu Kumar, Nitin Kumar Madheshiya, Shantanu Sarkar, Siddhant Sengar, students of the department of Computer Science and Engineering, Dr. B. C. Roy Engineering College, Durgapur, has successfully completed the project entitled Text To Speech as partial fulfillments for the award of the degree of B.Tech in Computer Science and Engineering.

This report is a bonafide piece of work done her and has not been submitted elsewhere for the award of any other degree.

____

INTERNAL EXAMINER

____

EXTERNAL EXAMINER

Project Name: Text To Speech

Project: **PR/CSE/14**

Team Report

Group PR/CSE/14

Title: Text To Speech

Basic Objective: To identify essential but unresolved requirements of TTS application and prototype implementation of those requirements.

Team Members:

Sl. No	Name	University Roll No	Signature
1.	SIDDHANT SENGAR	12000112097	
2.	BALENDU KUMAR	12000112030	
3.	SHANTANU SARKAR	12000112090	
4.	NITIN KUMAR MADHESHIYA	12000112065	

Mentors’ Signature:

ABSTRACT

The problem for developing a TTS (text-to-speech) is a very active field of research. As the Human-Computer Interfaces (HCI) come of age, the need for a more ergonomic and natural interface than the current one (keyboard, mouse, etc.) is being constantly felt. Talking of natural interfaces, what comes to mind, is sound (speech) and sight (vision). These form the basis of many intelligent systems research like robotics. Moreover, speech can also serve as an excellent interface for sightless people, or people with motor neuron disorders.

In this dissertation we attempt at developing a TTS system for English Language. Although the task of building very high quality, unlimited vocabulary text-to-speech (TTS) system is still a difficult one, with many open research questions, we believe the building of reasonable quality voices for many tasks can serve our needs. Here we have worked with English, the most commonly spoken language. We hope to easily extend the system to other languages, since there are a lot of underlying similarities between various languages. English language being highly phonetic, result in simple letter-to-sound rules. We used the standard concatenative synthesis. The main problem faced by us was to make the synthesized speech sound natural. We investigated the reasons for the mechanical sounding speech and developed different synthesis models to overcome some of those problems. Moreover, we implemented some standard and also novel intonation and duration modification algorithms, which can be incorporated into the TTS at a later stage. Our main achievement was reasonably legible speech with an unlimited vocabulary. The following thesis presents a brief overview of the main text-to-speech synthesis problem and its sub problems, and the initial work done in building a TTS for English.

Table of Contents

1. INTRODUCTION

Speech is the primary means of communication between people. Speech synthesis has been under development for several decades from now. Recent progress in speech synthesis has produced synthesizers with very high intelligibility.

1.1 SPEECH PROCESSING

In the year 1960, the world first time witnessed the idea of a talking computer. It was demonstrated in a movie theatre through a space odyssey with two astronauts and a computer. This computer was named HAL. HAL could not only speak but 88was also friendly and understanding. Before HAL, an actor speaking as a computer deliberately created a stylized, mechanical, "robotic" voice. That mechanical sound was the viewer's perception that a computer was speaking. However, HAL presented the possibility that future computers would speak and function like human beings. For most of the people, who were present in the demonstration of HAL, a computer was something out of science fiction. The way we interact with computers today - by typing on a keyboard to input information and receiving responses on a video screen - was just being designed at that time. With the invention of new technologies such as speech synthesis and speech recognition, we are now moving into an era of more effective human-computer interaction.

1.1.1 SPEECH TECHNOLOGY

Speech technology consists of the following two key components:

- Speech synthesis – Speech synthesis can be described in the simple words as computers speaking to people. This mainly requires computers to understand the language speaking rules. TTS synthesizers belong to this category.

- Speech recognition – Speech recognition can be considered as people speaking to computers. This requires computers to understand the speech. SPEECH-TO-TEXT systems belong to this category. In the present work, the speech synthesis component of speech technology has been considered.

1.1.2 SPEECH SYNTHESIS

Speech synthesis can also be considered as the generation of an acoustic speech signal by a computer. The application areas of speech synthesis may include:

- TTS synthesizers like e-mail or news readers, etc.
- Dialogue systems, for example, enquiry for train schedule information or information about flight reservation.
- Automatic translation (speech-to-speech) systems.
- Concept/content-to-speech (CTS), for example, weather forecasting.

1.1.3 TTS SYNTHESIZER

A Text-To-Speech (TTS) synthesizer is a computer-based system that should be able to read any text aloud, whether it was directly introduced in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. As such, the process of TTS conversion allows the transformation of a string of phonetic and prosodic symbols into a synthetic speech signal. The quality of the result produced by a TTS synthesizer is a function of the quality of the string, as well as of the quality of the generation process.

1.1.4 COMPONENTS OF A TTS SYNTHESIZER

As depicted in Fig. 1.1, a TTS synthesizer is composed of two parts:

- A front-end that takes input in the form of text and outputs a symbolic linguistic representation.
- A back-end that takes the symbolic linguistic representation as input and outputs the synthesized speech in waveform. These two phases are also called as high-level synthesis phase and low-level synthesis phase, respectively.

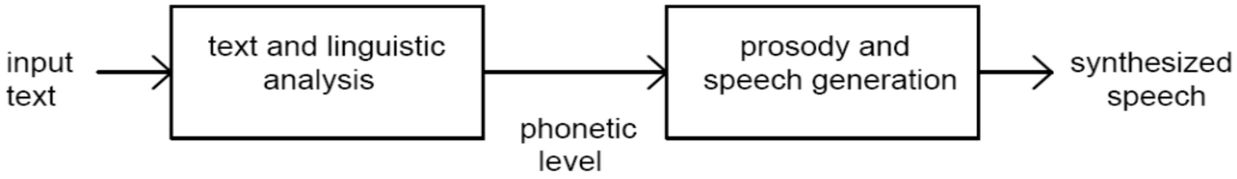

Fig. 1.1: Black-box view of a TTS synthesizer

1.1.5 TTS SYNTHESIZER VS. TALKING MACHINE/VOICE RESPONSE SYSTEMS

A TTS synthesizer differs from any other type of “talking machine” in the sense that it should be able to automatically produce “new sentences”, thus, it should also be able to read unrestricted text [Allen, 1976]. It is also important to note that a TTS synthesizer differs from any of the so-called voice response systems (such as announcements in a train etc.) in the sense that the vocabulary in a TTS synthesizer is not limited. The set consisting of utterance types is also not limited in TTS synthesizer.

1.1.6 QUALITY OF A TTS SYNTHESIZER

Usually, two quality criteria are proposed for deciding the quality of a TTS synthesizer.

- **Intelligibility** – it refers to how easily the output can be understood. It can be measured by taking into account several kinds of units (phonemes, syllables, words, phrases, etc.).
- **Naturalness or pleasantness** – it refers to how much the output sounds like the speech of a real person. Naturalness may be related to the concept of realism in the field of image synthesis: the goal is not to restitute the reality but to suggest it. Thus, listening to a synthetic voice must allow the listener to attribute this voice to some pseudo-speaker and to perceive some kind of expressivities as well as some indices characterizing the speaking style and the particular situation of elocution. For this purpose, the corresponding extra-linguistic information must be supplied to the synthesizer. Most of the existing TTS synthesizers produce an acceptable level of intelligibility, but the naturalness dimension, the ability to control expressivities, speech style and pseudo-speaker identity are still now the areas of

concern and need improvements. However, users’ demands vary to a large extent according to the field of application: general public applications such as telephonic information retrieval need maximal realism and naturalness, whereas some applications involving professionals (process or vehicle control) or highly motivated persons (visually impaired, applications in hostile environments) demand intelligibility with the highest priority.

1.2 APPLICATIONS OF SYNTHESIZED SPEECH

Synthetic speech can be used in a number of applications. Communication aids have developed from low quality talking calculators to modern 3D applications, such as talking heads. The implementation method varies from one application to the other. In some cases, such as announcement or warning systems, unrestricted vocabulary is not necessary and the best result is usually achieved with some simple messaging system. On the other hand, some applications, such as reading machines for the blind or electronic-mail readers, require unlimited vocabulary and a TTS synthesizer is needed.

The application field of synthetic speech is expanding fast whilst the quality of TTS synthesizers is also increasing steadily. Speech synthesizers are also becoming more affordable for common customers, which makes these synthesizers more suitable for everyday use. For example, better availability of TTS synthesizers may increase employing possibilities for people with communication deficiencies.

1.2.1 APPLICATIONS FOR THE BLIND – MOTIVATION FOR THE CURRENT WORK

Probably the most important and useful application field in speech synthesis is the reading and communication aids for the blind persons. Before the usage of synthesized speech, specific audio books were used where the content of the book was read into audio tape. It is clear that making such spoken copy of any large book takes several months and is very expensive.

These days, the synthesizers are mostly software based. As such with scanner and OCR system, it is easy to construct a reading machine for any computer environment with tolerable expenses. Speech synthesis is currently used to read www-pages or other forms of media with normal personal computer.

1.2.2 APPLICATIONS FOR THE DEAFENED AND VOCALLY HANDICAPPED

Voice handicaps originate in mental or motor/sensation disorders. Machines can be an invaluable support in the latter case. With the help of an especially designed keyboard and a fast sentence-assembling program, synthetic speech can be produced in a few seconds to remedy these impediments.

1.2.3 EDUCATIONAL APPLICATIONS

Synthesized speech can also be used in many educational situations. A computer with speech synthesizer can teach 24 hours a day and 365 days a year. It can be programmed for special tasks like spelling and pronunciation teaching for different languages. TTS synthesis coupled with a Computer Aided Learning system can provide a helpful tool to learn a new language. Synthesized speech can also be used with interactive educational applications.

Especially, with people who are impaired to read, speech synthesis may be very helpful because especially some children may feel themselves very embarrassing when they have to be helped by a teacher. It is also almost impossible to learn write and read without spoken help. With proper computer software, unsupervised training for these problems is easy and inexpensive to arrange. A speech synthesizer connected with word processor is also a helpful aid to proof reading. Many users find it easier to detect grammatical and stylistic problems when listening than reading. Normal misspellings are also easier to detect.

1.2.4 APPLICATIONS FOR TELECOMMUNICATIONS AND MULTIMEDIA

The newest applications in speech synthesis are in the area of multimedia. Synthesized speech has been used for decades in all kind of telephone enquiry systems, but the quality has been far from good for common customers. Today, the quality has reached the level that normal customers are adopting it for everyday use. Texts might range from simple messages, such as local cultural events to huge databases, which can hardly be read and stored as digitized speech. Synthetic speech is key factor in voice mail systems. This is a well-known fact that electronic mails have become very usual in last few years. However, it is not possible to read the mails without proper connectivity. This type of situation may exist at various places, e.g., a person traveling in the plane. With synthetic speech, e-mail messages may be listened to via normal telephone line. Synthesized speech may also be used to speak out short text messages (SMS) in mobile phones. For totally interactive multimedia applications an automatic speech recognition system is also needed. The automatic recognition of fluent speech is still far away, but the quality of current synthesizers is at least so good that it can be used to give some control commands, such as yes/no, on/off, or ok/cancel.

1.2.5 FUNDAMENTAL AND APPLIED RESEARCH TTS

Synthesizers possess a very peculiar feature, which makes them wonderful laboratory tools for linguists. These are completely under control, so that repeated experiences provide identical results (as is hardly the case with human beings). Consequently, they allow investigating the efficiency of into native and rhythmic models. A particular type of TTS synthesizer, that is based on a description of the vocal tract through its resonant frequencies (its formants) and denoted as formant synthesizer, has also been extensively used by phoneticians to study speech in terms of acoustical rules.

1.2.6 OTHER APPLICATIONS AND FUTURE DIRECTIONS

In principle, speech synthesis may be used in all kind of human-machine interactions. For example, in warning and alarm systems synthesized speech may be used to give more accurate information of the current situation. Using speech instead of warning lights or buzzers gives an opportunity to reach the warning signal for example from a different room. Speech synthesizer may also be used to receive some desktop messages from a computer, such as printer activity or received e-mail. 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. If it is possible to recognize speech, transcribe it into UNICODE/ASCII string, and then resynthesize it back to speech, a large amount of transmission capacity may be saved. With talking mobile phones, it is possible to increase the usability considerably for example with visually impaired users or in situations where it is difficult or even dangerous to try to reach the visual information. It is obvious that it is less dangerous to listen than to read the output from mobile phone for example when driving a car. During last few decades the communication aids have been developed from talking calculators to modern three-dimensional audio-visual applications. The application field for speech synthesis is becoming wider, which also brings more funds into research and development areas.

2. BASIC REQUIREMENTS OF THE PROJECT

The project is about the development of the Speech Synthesis System that could be used input text to analyse, synthesize and generate the output in the form of the audible sound.

Speech technology primarily comprises of two speech engines called as Speech Synthesis and Speech Recognition. Speech recognition is the mechanisms by which human speech is translated into text that an application would understand, whereas speech synthesis is the process of translating text into human understandable speech. It is also known as Text-to-Speech (TTS) conversion. The system model only uses the speech synthesis engine to convert text into speech. Hence, Speech recognition engine is not an issue.

During speech synthesis, first analysis of the text takes place that comprises of several steps like prosody analysis, structure analysis and text-to-phoneme conversion. In prosody analysis, speech attributes like the pitch, pausing, timing, peaking rate and others are concerned. After completion of these analysis the text’s prosody and structural information is gathered for conversion of it into phonetic or any linguistic description. Further, the prosody and phonetic information are used for the generation of speech waveforms using any of the two ways. One is to use words of pre-recorded human speech and concatenate them for full speech generation and other is to use signal processing techniques that are based on the mechanism of the phonemes sound and how they are affected by the prosody.

3. ARCHITECTURE ASSUMPTION: DETAILED OVERVIEW

3.1. How does a computer read?

This apparently difficult question can be answered by looking into the architecture of a typical TTS synthesizer. This section discusses in more detail the two components, namely, Natural Language Processing (NLP) component and Digital Signal Processing (DSP) component (Fig. 3.1) of a typical TTS synthesizer.

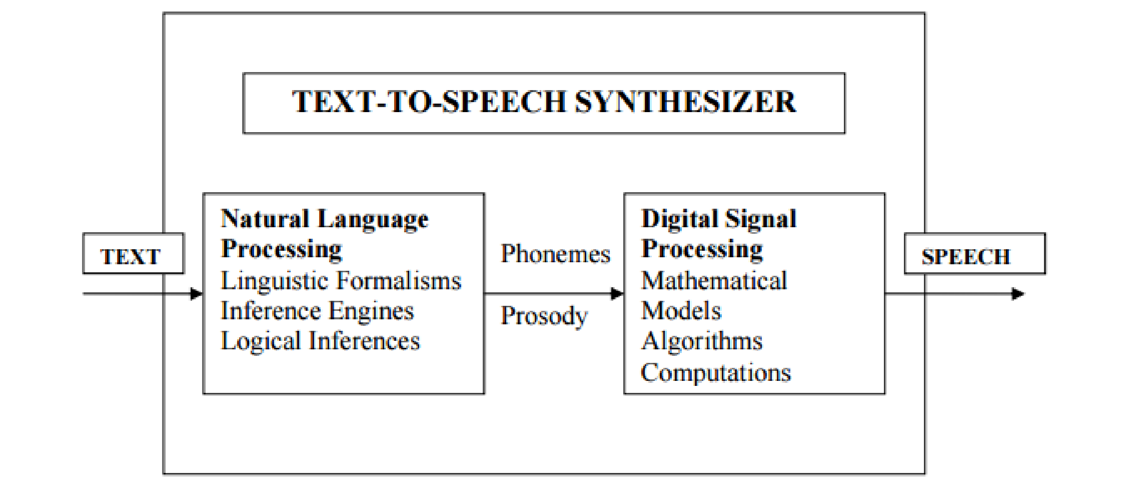

Fig. 3.1: Basic components of a TTS synthesizer

3.2 NLP COMPONENT

A general NLP module can be seen as consisting of three major subcomponents (Fig. 3.2).

Component 1: TEXT ANALYSIS

The text analyser consists of the following four modules:

- A pre-processing module, which organizes the input sentences into manageable lists of words. It identifies numbers, abbreviations, acronyms and idiomatic and transforms them into full text where so ever needed.

- A morphological analysis module, the task of which is to propose all possible part of speech categories for each word taken individually, on the basis of their spelling.

- The contextual analysis module considers words in their context, which allows it to reduce the list of their possible part of speech categories to a very restricted number of highly probable hypotheses, given the corresponding possible parts of speech of neighbouring words.

- Finally, a syntactic-prosodic parser, which examines the remaining search space and finds the text structure (i.e., its organization into clause and phrase-like constituents) which more closely relates to its expected prosodic realization.

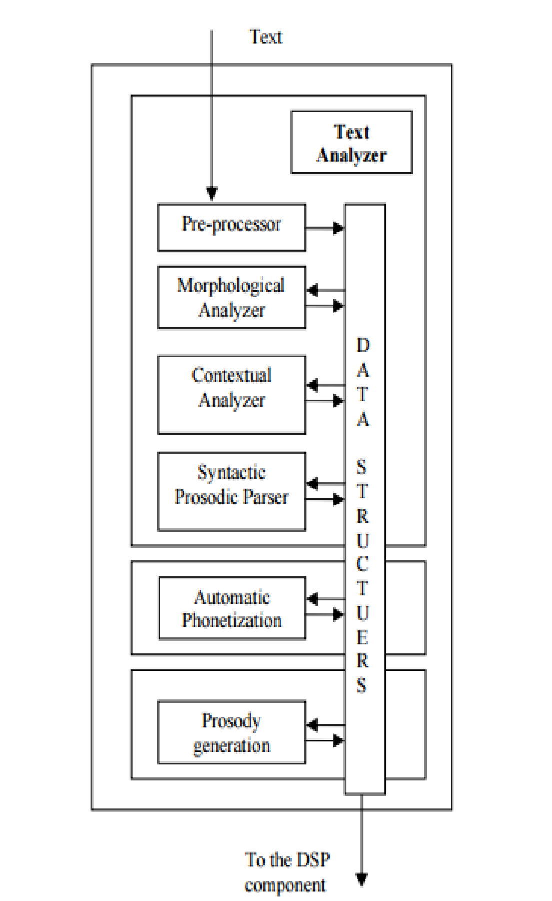

Fig. 3.2: NLP Component

Component 2: AUTOMATIC PHONETIZATION

The Automatic Phonetization module (also known as Letter-To-Sound (LTS) module) is responsible for the automatic determination of the phonetic transcription of the incoming text. Pronunciation dictionaries may not always help in this module due to the following facts:

- The pronunciation dictionaries refer to word roots only. These do not explicitly account for morphological variations (i.e. plural, feminine, conjugations, especially for highly inflected languages).

- Words embedded into sentences are not pronounced as if they were isolated. It can also be noted that not all words can be found in a phonetic dictionary. The pronunciation of new words and of many proper names has to be deduced from the one of already known words.

The two popular ways of implementing an Automatic Phonetization module are:

1. Dictionary-based solutions that consist of storing a maximum of phonological knowledge into a lexicon.

2. Rule-based transcription systems that transfer most of the phonological competence of dictionaries into a set of letter-to-sound (or grapheme-to phoneme) rules. This time, only those words that are pronounced in such a particular way that they constitute a rule on their own are stored in an exceptions dictionary.

Component 3: PROSODY GENERATION

Prosodic features consist of pitch, duration, and stress over the time. With a good control over these features, gender, age, emotions, and other features can be incorporated in the speech and very natural sounding speech can be modelled. However, almost everything seems to have an effect on prosodic features of natural speech and it makes accurate modelling very difficult (Fig. 3.3).

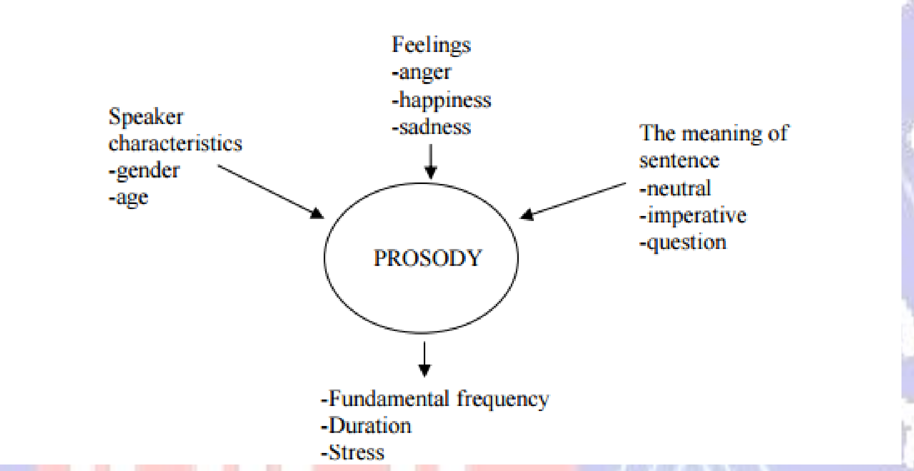

Fig. 3.3: Prosodic dependencies

Prosodic features can be divided into several levels such as syllable, word, and phrase level. For example, at word level vowels are more intense than consonants. At phrase level correct prosody is more difficult to produce than at the word level. The three features of the prosody are described below in brief.

Pitch: The pitch pattern or fundamental frequency over a sentence (intonation) in natural speech is a combination of many factors. The pitch contour depends on the meaning of the sentence. For example, in normal speech the pitch slightly decreases towards the end of the sentence and when the sentence is in a question form, the pitch pattern will raise to the end of sentence. In the end of sentence, there may also be a continuous rise which indicates that there is more speech to come. A raise or fall in fundamental frequency can also indicate a stressed syllable. Finally, the pitch contour is also affected by gender, physical and emotional state, and attitude of the speaker.

Duration: The duration or time characteristics can also be investigated at several levels from phoneme (segmental) durations to sentence level timing, speaking rate, and rhythm. The segmental duration is determined by a set of rules to determine correct timing. Usually, some inherent duration for phoneme is modified by rules between maximum and minimum durations. In general, the phoneme duration differs due to neighbouring phonemes. At sentence level, the speech rate, rhythm, and correct placing of pauses for correct phrase boundaries are important.

Intensity: The intensity pattern is perceived as a loudness of speech over the time. At syllable level, vowels are usually more intense than consonants and at a phrase level, syllables at the end of an utterance can become weaker in intensity. The intensity pattern in speech is highly related with fundamental frequency. The intensity of a voiced sound goes up in proportion to fundamental frequency.

3.3 DSP COMPONENT

The DSP component is also called as the synthesizer component. Different TTS synthesizers can be classified according to the type of synthesis technique that is used to synthesize the speech. The methods are usually classified into three groups:

Articulatory synthesis – this attempts to model the human speech production system directly and is thus, potentially the most satisfying method to produce high quality synthetic speech. Articulatory synthesizer generates speech by mathematically modelling the movement of articulators, e.g., lips, tongue, and jaws.

Formant synthesis - this models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model. Thus speech is generated by an acoustic-phonetic production model, based on formant of the vocal tract. The acoustic-phonetic parameters such as energy, pitch and resonance (formant) frequencies associated with speech and heuristic rules are used to derive the model.

Concatenative synthesis -this uses different length pre-recorded samples derived from natural speech. Here, speech is generated by combining splices of pre-recorded natural speech.

The articulatory and formant synthesis are also classified as rule-based synthesis methods whereas the concatenative technique falls under database driven synthesis method. The formant and concatenative methods are the most commonly used in present synthesizers. The formant synthesis was dominant for long time, but these days, the concatenative method is becoming more and more popular. The articulatory method is still too complicated for high quality implementations, but may arise as a potential method in the future.

Some other synthesis methods are:

- **Hybrid synthesis** marries aspects of formant and concatenative synthesis to minimize the acoustic glitches when speech segments are concatenated.

- **HMM-based synthesis** is a synthesis method based on Hidden Markov Models (HMMs). In this type of synthesizer, speech frequency spectrum (vocal tract), Fundamental frequency (vocal source), and duration (prosody) are modelled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves based on Maximum likelihood criterion.

- **Sinusoidal Model based Synthesis** - Sinusoidal models are based on a well-known assumption that the speech signal can be represented as a sum of sine waves with time varying amplitude and frequencies.

- **Linear predictive methods** - Linear predictive methods are originally designed for speech coding systems, but may also be used in speech synthesis. In fact, the first speech synthesizers were developed from speech coders. Like formant synthesis, the basic LPC is based on the source-filter-model of speech described. The digital filter coefficients are estimated automatically from a frame of natural speech.

The presented work uses concatenative synthesis with syllable-like units for the development of TTS synthesizer. So, in the next section, details only about concatenative synthesis are given.

3.2.1 CONCATENATIVE SYNTHESIS

In the last decade, there has been a significant trend for development of speech synthesizers using Concatenative Synthesis techniques. There are a number of different methodologies for Concatenative Synthesis such as TDPSOLA, PSOLA, MBROLA and Epoch Synchronous Non over Lapping Add (ESNOLA).

There are three main subtypes of concatenative synthesis:

- **Unit selection synthesis**: This type of synthesis uses large speech databases (more than one hour of recorded speech). During database creation, each recorded utterance is segmented into some or all of the following: individual phones, syllables, morphemes, words, phrases, and sentences. Typically, the division into segments is done using a specially modified speech recognizer set to a "forced alignment" mode with some hand correction afterward, using visual representations such as the waveform and spectrogram. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighbouring phones.

- At runtime, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection). This process is typically achieved using a specially-weighted decision tree. Unit selection gives the greatest naturalness due to the fact that it does not apply a large amount of digital signal processing to the recorded speech, which often makes recorded speech sound less natural, although some synthesizers may use a small amount of signal processing at the point of concatenation to smooth the waveform.

- **Diphone synthesis**: It uses a minimal speech database containing all the Diphones (sound-to-sound transitions) occurring in a given language. The number of diphones depends on the phono tactics of the language: Spanish has about 800 diphones and German has about 2500 diphones. In diphone synthesis, only one example of each diphone is contained in the speech database. At runtime, the target prosody of a sentence is superimposed on these minimal units by means of digital signal processing techniques such as Linear predictive coding, PSOLA or MBROLA.
- The quality of the resulting speech is generally not as good as that from unit selection but more natural-sounding than the output of formant synthesizers. Diphone synthesis suffers from the sonic glitches of concatenative synthesis and the robotic-sounding nature of formant synthesis, and has few of the advantages of either approach other than small size. As such, its use in commercial applications is declining, although it continues to be used in research because there are a number of freely available implementations.

- **Domain-specific synthesis**: It concatenates pre-recorded words and phrases to create complete utterances. It is used in applications where the variety of texts the synthesizer will output is limited to a particular domain, like trains schedule announcements or weather reports. This technology is very simple to implement, and has been in commercial use for a long time: this is the technology used by gadgets like talking clocks and calculators.

- The naturalness of these synthesizers can potentially be very high because the variety of sentence types is limited and closely matches the prosody and intonation of the original recordings.

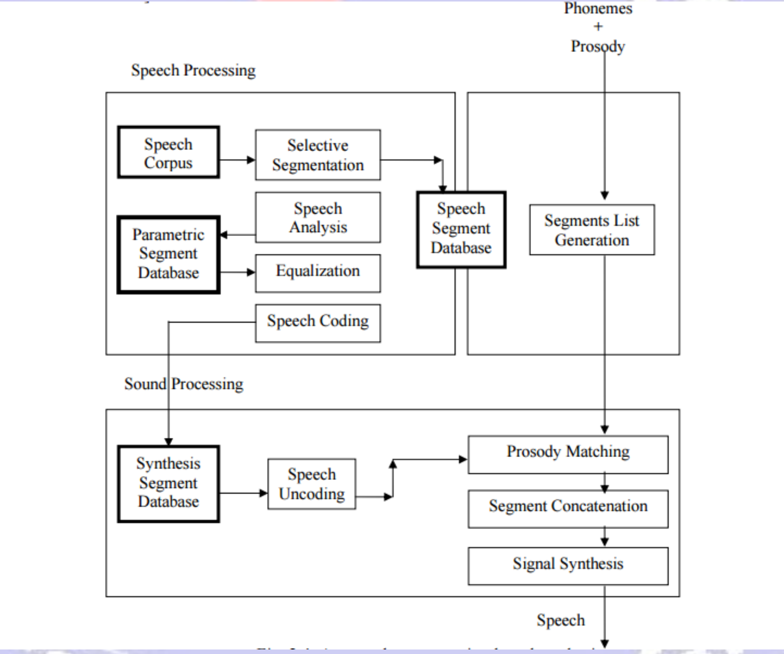

Fig. 3.4: A general concatenation-based synthesizer

4. SYSTEM DESIGN

Figure 4.1 introduces the functional diagram of a very general TTS synthesizer. As for human reading, it comprises a Natural Language Processing module (NLP), capable of producing a phonetic transcription of the text read, together with the desired intonation and rhythm (often termed as prosody), and a Digital Signal Processing module (DSP), which transforms the symbolic information it receives into speech. But the formalisms and algorithms applied often manage, thanks to a judicious use of mathematical and linguistic knowledge of developers, to short-circuit certain processing steps. This is occasionally achieved at the expense of some restrictions on the text to pronounce, or results in some reduction of the "emotional dynamics" of the synthetic voice (at least in comparison with human performances), but it generally allows to solve the problem in real time with limited memory requirements.

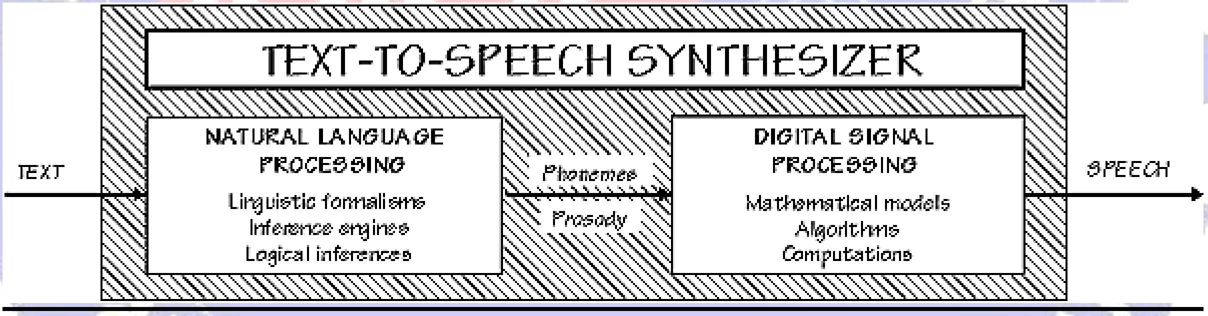

Fig. 4.1 General System Design of TTS System

5. TOOLS USED FOR DEVELOPMENT OF A TTS SYNTHESIZER

The tools available for developing a TTS synthesizer include speech API's provided by different vendors, and different markup languages. There exist many different APIs for speech output but Responsivevoice.js is a remarkable tool for designing a Text to Speech engine with multiple voice accents.

5.1 ResponsiveVoice.JS API

ResponsiveVoice is a HTML5-based Text-To-Speech library designed to add voice features to web sites and apps across all smartphone, tablet and desktop devices. It supports 51 languages through 168 voices, no dependencies and weighs just 14kb.

HTML5 introduces the Speech API for Speech Synthesis and Speech Recognition. This is the easiest way to use the spoken word in your app or website. Speech Synthesis or more commonly known as Text To Speech (TTS) is now available in most modern browsers. Gone are the days of waiting for Text To Speech engines to render MP3 audio files from text and then download them from servers. Today the browser can instantly speak text on the client side and with quite reasonable quality.

ResponsiveVoice JS defines a selection of smart Voice profiles that know which voice to use for the users device in order to create a consistent experience no matter which browser or device the speech is being spoken on.

By choosing one ResponsiveVoice the closest voice is chosen on

- iOS (Safari & Chrome)
- Android (Chrome, Including across the popular Text To Speech engines Ivona, Acapela, Samsung)
- Windows (Chrome Desktop)
- Mac OSX (Safari & Chrome)

5.1.1 How it Works

INCLUDE THE JS FILE IN YOUR PAGE

```
<script src="http://code.responsivevoice.org/responsivevoice.js"></script>
```

SPEAK(STRING TEXT, [STRING VOICE], [OBJECT PARAMETERS])

Starts speaking the text in a given voice.

Parameters

text: String

The text to be spoken.

voice: String

Defaults to “UK English Female”. Choose from the available ResponsiveVoices.

parameters: Object

Used to add optional pitch (range 0 to 2), rate (range 0 to 1.5), volume (range 0 to 1) and callbacks.

Pitch, rate and volume may not affect audio on some browser combinations, older versions of Chrome on Windows for example.

responsiveVoice.speak("hello world");

responsiveVoice.speak("hello world", "UK English Male");

responsiveVoice.speak("hello world", "UK English Male", {pitch: 2});

responsiveVoice.speak("hello world", "UK English Male", {rate: 1.5});

responsiveVoice.speak("hello world", "UK English Male", {volume: 1});

responsiveVoice.speak("hello world", "UK English Male", {onstart: StartCallback, onend: EndCallback});

Returns: true/false

CANCEL( )

Stops playing the speech.

```
responsiveVoice.cancel();
```

VOICESUPPORT( )

Checks if browser supports native TTS

```
if(responsiveVoice.voiceSupport())
{
responsiveVoice.speak("hello world");
}
```

Returns: *true/false*

GETVOICES( )

```
var voicelist = responsiveVoice.getVoices();
```

Returns: a list of available voices

SETDEFAULTVOICE( )

```
responsiveVoice.setDefaultVoice("US English Female");
```

ISPLAYING( )

Detects if native TTS or TTS audio element is producing output.

```
if(responsiveVoice.isPlaying())
{
  console.log("I hope you are listening");
}
```

Returns: true/false

PAUSE() AND RESUME( )

Pauses/Resumes speech

```
responsiveVoice.pause();
responsiveVoice.resume();
```

5.1.2 Special Features

ResponsiveVoice JS also takes care of a number of hindrances from the various implementations of HTML5 Speech API across browsers and operating systems.

- Chrome desktop has a limit on the number of characters it can speak, under the hood ResponsiveVoice JS automatically chunks text into acceptable blocks
- Chrome desktop will not speak unless initialised after page load, ResponsiveVoice JS resolves this
- iOS Safari & Chrome require timing delays between speech API calls, ResponsiveVoice JS resolves this
- iOS TTS can’t be triggered without a direct user interaction, ResponsiveVoice JS resolves this
- Internet Explorer speech rate is slower, ResponsiveVoice JS resolves this

With large blocks of text ResponsiveVoice splits up the text into chunks, with preference given to splitting at the end of sentences. Preference is given to splitting at full stop, question mark, colon or semi-colon after that split is performed by the nearest comma and falling back from that the nearest space between words.

5.1.3 Limitations of ResponsiveVoice API

There is a problem, each browser and device can have a different set of “Voices”. You can’t be sure of a consistent user experience when it comes to the spoken voice or accent. If you make a call to the speak API using the default voice it will sound very different on different users devices and browsers. In some cases you won’t even know if the user will get a male or female voice.

Although, you make a direct call to the speak API and choose a specific voice like “Google UK Female”, if a user is browsing on iOS with Safari the voice will not be available.

6.CONCLUSION

It is quite clear that there is still very long way to go before text-to-speech synthesis, especially high-level synthesis, is fully acceptable. However, the development is going forward steadily and in the long run the technology seems to make progress faster than we can imagine. Thus, when developing a speech synthesis system, we may use almost all resources available, because in few years' todays high resources are available in every personal computer. Regardless how fast the development process will be, speech synthesis, whenever used in low-cost calculators or state-of-the-art multimedia solutions, has probably the most promising future. If speech recognition systems someday achieve a generally acceptable level, we may develop for example a communication system where the system may first analyze the speakers' voice and its characteristics, transmit only the character string with some control symbols, and finally synthesize the speech with individual sounding voice at the other end. After going through all the research part, the current work has been aimed to add more features to this TTS synthesizer apart from the existing one because project development is a never ending process and there is always some scope for addition and improvement.

7.REFERENCES

1. [Allen, 1976] Allen J (1976) Synthesis of speech from unrestricted text. IEEE Journal, Vol.64, Issue 4, pp 432-42

2. [Allen, 1987] Allen J, Hunnicutt S, Klatt D (1987) From text-to-speech: the MITalk system**.** Cambridge University Press, Inc.

3. [Banat, 1990] Banat Karima (Kuwait University), El-Imam Yousif A (IBM Kuwait Scientific Center) (1990) Text-to-speech conversion on a personal computer. IEEE Micro, Vol. 10, Issue 4, pp 62-74

4. [Black et al., 2001] User Manual for the Festival Speech Synthesis System, version 1.4.3 http://fife.speech.cs.cmu.edu/festival/cstr/festival/1.4.3/

5. [Black et al., 2001] Black A, Taylor P, Caley R (2001) The Festival speech synthesis system: system documentation. University of Edinburgh http://www.cstr.ed.ac.uk/projects/festival/

6. [Basu, 2002] Basu A, Choudhury M (2002) A rule based schwa deletion algorithm for Hindi. In: Proceedings of international conference on knowledgebased computer systems, IKON 2002, December, pp 343-53

7. [Dhvani, 2001] Dhvani-TTS system for Indian Languages http://dhvani.sourceforge.net

8. [Dutoit, 1996] Dutoit T (1996) An Introduction to Text-to-Speech Synthesis, First edition, Kluwer Academic Publishers

9. [Dutoit, 1996] Dutoit T (1996) High-quality text-to-speech synthesis: an overview. Journal of Electrical & Electronics Engineering, Australia: Special Issue on Speech Recognition and Synthesis, vol. 17, pp 25-37

10. [El-Imam, 1987] El-Imam Y A (1987) A personal computer-based speech analysis and synthesis system. IEEE Micro, Vol. 7, No. 3, June, pp 4-21

11. Responsive Voice Website. http://responsivevoice.org/