OCR BASED SPEECH SYNTHESIS SYSTEM USING LABVIEW

Kanchana V[1] and Abdul Shabeer H[2]

PG Scholar/VLSI Design,Department of Electronics And Communication Engg,Salem.

Research Dean/Assistant Professor,Department of Electronics And Communication Engg,Salem.

Abstract— The Optical Character Recognition is a mobile application. It uses smart mobile phones of windows platform. This paper combines the functionality of Optical Character Recognition and speech synthesizer. The objective is to develop user friendly application which performs image to speech conversion system using android phones. The OCR takes image as the input, gets text from that image and then converts it into speech. This system can be useful in various applications like banking, legal industry, other industries, and home and office automation. It mainly designed for people who are unable to read any type of text documents. In this paper, the character recognition method is presented by using OCR technology and windows phone with higher quality camera.

I. INTRODUCTION:

A. OCR:

OCR is the acronym for Optical Character Recognition. This technology allows to automatically recognizing characters through an optical mechanism. In case of human beings, our eyes are optical mechanism. The image seen by eyes is input for brain. The ability to understand these inputs varies in each person according to many factors [2]. OCR is a technology that functions like human ability of reading. Although OCR is not able to compete with human reading capabilities.

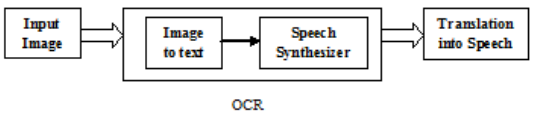


Fig.1.1 Block Diagram of OCR

OCR can recognize both handwritten and printed text. But the performance of OCR is directly dependant on quality of input documents.OCR is designed to process images that consist almost entirely of text, with very little non-text clutter obtain from picture captured by mobile camera. This application is for the Android mobile operating system that combines Google's open-source OCR engine, Tesseract, text recognition OCR engine [5]. Google's language translation service, and the Android operating system's text-to-speech synthesizer to allow users to take photographs of text using a camera phone and have the text read aloud. Most of the character recognition program will be recognized through the input image with a scanner or a digital camera and computer software. There is a problem in the spatial size of the computer and scanner. If you do not have a scanner and a digital camera, a hardware problem occurs. In order to overcome the limitations of computer occupying a large space, character recognition system based on android phone is proposed [4].

OCR is a technology that enables you to convert different types of documents such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data. Images captured by a digital camera differ from scanned documents or image. They often have defects such as distortion at the edges and dimmed light, making it difficult for most OCR applications, to correctly recognize the text. We have chosen Tesseract because of widespread approbation, its extensibility and flexibility, its community of active developers, and the fact that it "just works" out of the box. To perform the character recognition, our application has to go through three important steps. The first is Segmentation, i.e., given a binary input image, to identify the individual glyphs (basic units representing one or more characters, usually contiguous). The second step is feature extraction, i.e., to compute from each

glyph a vector of numbers that will serve as input features for an ANN [3]. This step is the most difficult in the sense that there is no obvious way to obtain these features. The final task is classification.

II.TEXT TO SPEECH:

A text to speech (TTS) synthesizer is a system that can read text aloud automatically, which is extracted from Optical Character Recognition (OCR). A speech synthesizer can be implemented by both hardware and software. Speech synthesis is the artificial production of human speech [9]. A computer system used for this purpose is called a speech synthesizer. A text-to-speech (TTS) system converts normal language text into speech. A synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

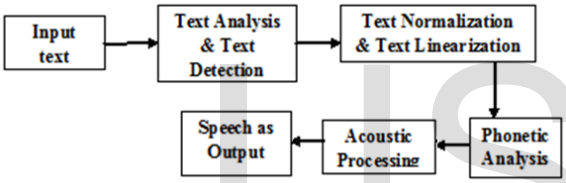


Fig.1.2 Text to Speech Conversion

a) Text Analysis & Detection:

The Text Analysis is part of preprocessing. It analyzes the input text and organizes into manageable list of words. Then it transforms them into full text. Text detection localizes the text areas from printed documents.

b) Text Normalization & Linearization:

Text Normalization is the transformation of text to pronounceable form. Text normalization is often performed before text is processed in some way, such as generating synthesized speech or automated language translation. The main objective of this process is to identify punctuation marks and pauses between words. Usually the text normalization process is done for converting all letters of lowercase or upper case, to remove punctuations, accent marks , stopwords or "too common words "and other diacritics from letters.

c) Phonetic Analysis:

It provides phonetic alphabets. The grapheme to phoneme conversion is done. It is actually a conversion of orthographical symbols into phonological symbols.

d) Acoustic Processing:

It performs formant synthesis. It works intelligently and thus does not require any kind of database of speech samples. For speak out the text, it uses voice characteristics of a person.

III.Speech synthesis:

Speech synthesis is the artificial production of human speech. Synthesizing is the very effective process of generating speech waveforms using machines based on the phonetical transcription of the message. Recent progress in speech synthesis has produced synthesizers with very high intelligibility but the sound quality and naturalness still remains a major problem.

IV.Phonetics and Theory of Speech Production:

Speech processing and language technology contains lots of special concepts and terminology. To understand how different speech synthesis and analysis methods work one must have some knowledge of speech production, articulatory phonetics, and some other related terminology. The basic theories related to these topics are described below.

V.Representation And Analysis of Speech Signals:

Continuous speech is a set of complicated audio signals which makes producing them artificially difficult. Speech signals are usually considered as voiced or unvoiced, but in some cases they are something between these two. Voiced sounds consist of fundamental frequency (F0) and its harmonic components produced by vocal cords (vocal folds). The vocal tract modifies this excitation signal causing formant (pole) and sometimes antiformant (zero) frequencies [8]. Each formant frequency has also amplitude and bandwidth and it may be sometimes difficult to define some of these parameters correctly. The fundamental frequency and formant frequencies are probably the most important concepts in speech synthesis and also in speech processing in general.

With purely unvoiced sounds, there is no fundamental frequency in excitation signal and therefore no harmonic structure either and the excitation can be considered as white noise. The airflow is forced through a vocal tract constriction which can occur in several places between glottis and mouth. Some sounds are produced with complete stoppage of airflow followed by a sudden release, producing an impulsive turbulent excitation often followed by a more protracted turbulent excitation [9]. Unvoiced sounds are also usually more silent and less steady than voiced ones.

Speech signals of the three vowels (/a/ /i/ /u/) are presented in time- and frequency domain in Figure: 1.4. The fundamental frequency is about 100 Hz in all cases and the formant frequencies F1, F2, and F3 with vowel /a/ are approximately 600 Hz, 1000 Hz, and 2500 Hz respectively. With vowel /i/ the first three formants are 200 Hz, 2300 Hz, and 3000 Hz, and with /u/ 300 Hz, 600 Hz, and 2300 Hz. The harmonic structure of the excitation is also easy to perceive from frequency domain presentation.

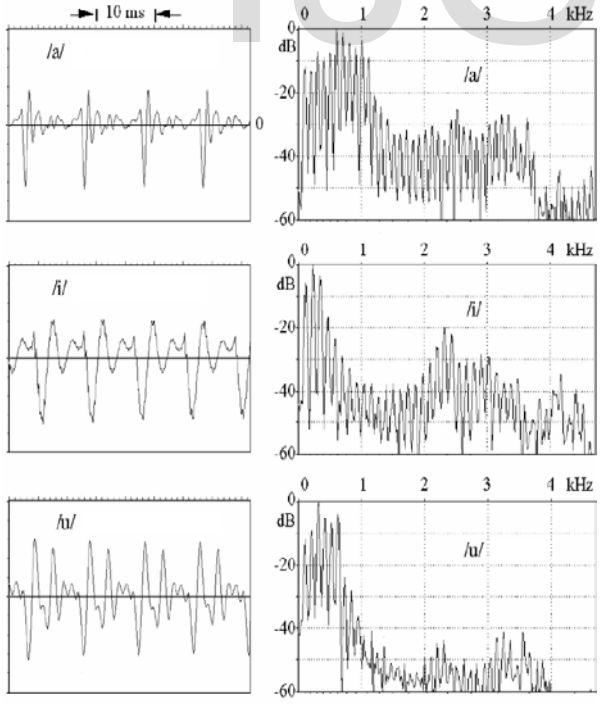


Figure 1.3 The time- and frequency-domain presentation of vowels /a/, /i/, and /u/.

For determining the fundamental frequency or pitch of speech, for example a method called cepstral analysis may be used [9]. Cepstrum is obtained by first windowing and making Discrete Fourier Transform (DFT) for the signal and then logaritmizing power spectrum and finally transforming it back to the time-domain by Inverse Discrete Fourier Transform (IDFT).

Fundamental frequency or intonation contour over the sentence is important for correct prosody and natural sounding speech. The different contours are usually analysed from natural speech in specific situations and with specific speaker characteristics and then applied to rules to generate the synthetic speech. The fundamental frequency contour can be viewed as the composite set of hierarchical patterns shown in Figure 1.4. The overall contour is generated by the superposition of these patterns [10].

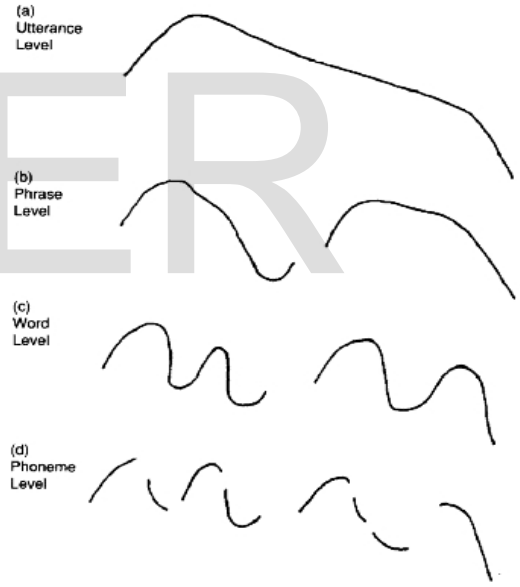


Figure 1.4 Hierarchical levels of fundamental frequency.

VI.Speech Production:

Human speech is produced by vocal organs presented in Figure 1.5. The main energy source is the lungs with the diaphragm. When speaking, the air flow is forced through the glottis between the vocal cords and the larynx to the three main cavities of the vocal tract, the pharynx and the oral and nasal cavities. From the oral and nasal cavities the air flow exits through the nose and mouth,

respectively. The V-shaped opening between the vocal cords, called the glottis, is the most important sound source in the vocal system. The vocal cords may act in several different ways during speech. The most important function is to modulate the air flow by rapidly opening and closing, causing buzzing sound from which vowels and voiced consonants are produced. The fundamental frequency of vibration depends on the mass and tension and is about 110 Hz, 200 Hz, and 300 Hz with men, women, and children, respectively. With stop consonants the vocal cords may act suddenly from a completely closed position, in which they cut the air flow completely, to totally open position producing a light cough or a glottal stop. On the other hand, with unvoiced consonants, such as /s/ or /f/, they may be completely open. An intermediate position may also occur with for example phonemes like /h/.

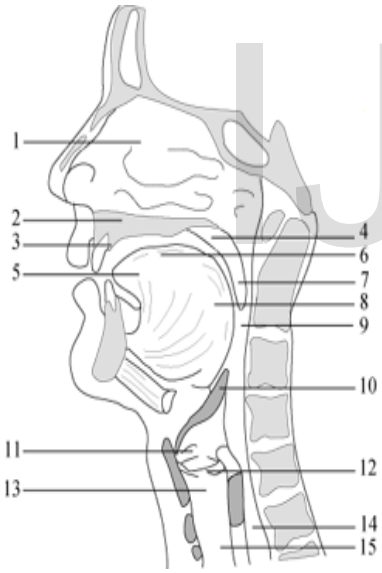


Figure 1.5 The human vocal organs. (1) Nasal cavity, (2) Hard palate, (3) Alveoral ridge, (4) Soft palate (Velum), (5) Tip of the tongue (Apex), (6) Dorsum, (7) Uvula, (8) Radix, (9) Pharynx, (10) Epiglottis, (11) False vocal cords, (12) Vocal cords, (13) Larynx, (14) Esophageus, and (15) Trachea.

The pharynx connects the larynx to the oral cavity. It has almost fixed dimensions, but its length may be changed slightly by raising or lowering the larynx at one end and the soft palate at the other end. The soft palate also isolates or connects the route from the nasal cavity to the pharynx. At the bottom of the pharynx are the epiglottis and false vocal cords to prevent food reaching the larynx and to isolate the esophagus acoustically from the vocal tract. The epiglottis, the false vocal cords and the vocal cords are closed during swallowing and open during normal breathing.

The oral cavity is one of the most important parts of the vocal tract. Its size, shape and acoustics can be varied by the movements of the palate, the tongue, the lips, the cheeks and the teeth. Especially the tongue is very flexible, the tip and the edges can be moved independently and the entire tongue can move forward, backward, up and down. The lips control the size and shape of the mouth opening through which speech sound is radiated. Unlike the oral cavity, the nasal cavity has fixed dimensions and shape. Its length is about 12 cm and volume 60 cm^3. The air stream to the nasal cavity is controlled by the soft palate.From technical point of view, the vocal system may be considered as a single acoustic tube between the glottis and mouth. Glottal excited vocal tract may be then approximated as a straight pipe closed at the vocal cords where the acoustical impedance Z_g= ¥ and open at the mouth (Z_m = 0) [11, 12]. In this case the volume-velocity transfer function of vocal tract is If l=17 cm, V(w) is infinite at frequencies F_i = 500, 1500, 2500, Hz which means resonances every 1 kHz starting at 500 Hz. If the length l is other than 17 cm, the frequencies F_i will be scaled by factor 17/l so the vocal tract may be approximated with two or three sections of tube where the areas of adjacent sections are quite different and resonances can be associated within individual cavities. Vowels can be approximated with a two-tube model presented on the left in Figure 1.8. For example, with vowel /a/ the narrower tube represents the pharynx opening into wider tube representing the oral cavity. If assumed that both tubes have an equal length of 8.5 cm, formants occur at twice the frequencies noted

earlier for a single tube. Due to acoustic coupling, formants do not approach each other by less than 200 Hz so formants F1 and F2 for /a/ are not both at 1000 Hz, but rather 900 Hz and 1100 Hz, respectively.

The excitation signal may be modeled with a two-mass model of the vocal cords which consists of two masses coupled with a spring and connected to the larynx by strings and dampers.

Phonetics:

In most languages the written text does not correspond to its pronunciation so that in order to describe correct pronunciation some kind of symbolic presentation is needed. Every language has a different phonetic alphabet and a different set of possible phonemes and their combinations. The number of phonetic symbols is between 20 and 60 in each language. A set of phonemes can be defined as the minimum number of symbols needed to describe every possible word in a language. In English there are about 40 phonemes. Due to complexity and different kind of definitions, the number of phonemes in English and most of the other languages cannot be defined exactly.

Phonemes are abstract units and their pronunciation depends on contextual effects, speaker's characteristics, and emotions. During continuous speech, the articulatory movements depend on the preceding and the following phonemes. The articulators are in different position depending on the preceding one and they are preparing to the following phoneme in advance. This causes some variations on how the individual phoneme is pronounced. These variations are called allophones which are the subset of phonemes and the effect is known as co-articulation. For example, a word lice contains a light /l/ and small contains a dark /l/. These l's are the same phoneme but different allophones and have different vocal tract configurations. Another reason why the phonetic representation is not perfect is that the speech signal is always continuous and phonetic notation is always discrete [8]. Different emotions and speaker characteristics are also impossible to describe with phonemes so the unit called phone is usually defined as an acoustic realization of a phoneme.

The phonetic alphabet is usually divided in two main categories, vowels and consonants. Vowels are always voiced sounds and they are produced with the vocal cords in vibration, while consonants may be either voiced or unvoiced. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyze and describe acoustically. Because consonants involve very rapid changes they are more difficult to synthesize properly. The articulatory phonetics in English and Finnish are described more closely in the next section of this chapter.

Some efforts to construct language-independent phonemic alphabets were made during last decades. One of the best known is perhaps IPA (International Phonetic Alphabets) which consists of a huge set of symbols for phonemes, suprasegmentals, tones/word accent contours, and diacritics. For example, there are over twenty symbols for only fricative consonants (IPA 1998). Complexity and the use of Greek symbols makes IPA alphabet quite unsuitable for computers which usually requires standard ASCII as input. Another such kind of phonetic set is SAMPA (Speech Assessment Methods Phonetic Alphabet) which is designed to map IPA symbols to 7-bit printable ASCII characters. In SAMPA system, the alphabets for each language are designed individually. Originally it covered European Communities languages, but the objective is to make it possible to produce a machine-readable phonetic transcription for every known human language. Alphabet known as Worldbet is another ASCII presentation for IPA symbols which is very similar to SAMPA. American linguists have developed the Arpabet phoneme alphabet to represent American English phonemes using normal ASCII characters. For example a phonetic representation in DECtalk system is based on IPA and Arpabet with some

modifications and additional characters. Few examples of different phonetic notations are given in Table 1.1.

Table: 1.1 Examples of different phonetic notations.

IPA	IPA-ASCII	SAMPA	DECtalk	Example
i	i	i:	iy	beet
I	I	I	ih	bit
ɛ	E	e	ey	bet
æ	&	{	ae	at
ə	@	@	ax	about
ʌ	V	V	ah	but

Several other phonetic representations and alphabets are used in present systems. For example MITalk uses a set of almost 60 two-character symbols for describing phonetic segments in it and it is quite common that synthesis systems use the alphabet of their own. There is still no single generally accepted phonetic alphabet.

VII.English Articulatory Phonetics:

Unlike in Finnish articulatory phonetics, discussed bellow, the number of phonetic symbols used in English varies by different kind of definitions. Usually there are about ten to fifteen vowels and about twenty to twenty-five consonants.

English vowels may be classified by the manner or place of articulation (front- back) and by the shape of the mouth (open-close). Main vowels in English and their classification are described in Figure 1.9 below. Sometimes also some diphthongs like /ou/ in tone or /ei/ in take are described separately.

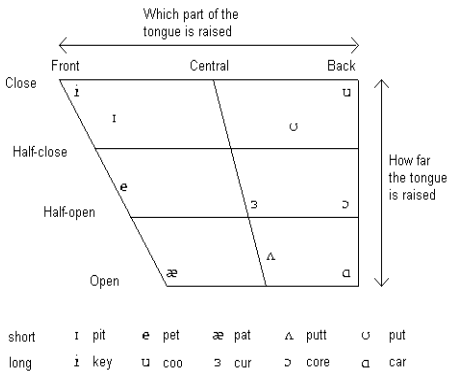


Figure 1.6 The classification of the main vowels in English.

English consonants may be classified by the manner of articulation as plosives, fricatives, nasals, liquids, and semivowels. Plosives are known also as stop consonants. Liquids and semivowels are also defined in some publications as approximants and laterals. Further classification may be made by the place of articulation as labials (lips), dentals (teeth), alveolars (gums), palatals (palate), velars (soft palate), glottal (glottis), and labiodentals (lips and teeth). Classification of English consonants is summarized in Figure 1.7.

place / manner	labial	labio-dental	dental	alveolar	palate-alveoral	palatal	velar	glottal
plosive	p b			t d			k g	
fricative		f v	q ð	s z	ʒ ʃ			h
nasal	m			n			ŋ	
liquid				r l				
semivowel	w					j		

Figure 1.7 Classifications of English Phonetics.

VIII.Finnish Articulatory Phonetics:

There are eight vowels in Finnish. These vowels can be divided into different categories depending how they are formulated: Front/back position of tongue, wideness/roundness of the constriction position, place of the tongue (high or low), and how open or close the mouth is during articulation.

Finnish consonants can be divided into the following categories depending on the place and the manner of articulation:

1.Plosives or stop consonants: /k, p, t, g, b, d/. The vocal tract is closed causing stop or attenuated sound. When the tract reopens, it causes noise-like, impulse-like or burst sound.

2.Fricatives: /f, h, s/. The vocal tract is constricted in some place so the turbulent air flow causes noise which is modified by the vocal tract resonances. Finnish fricatives are unvoiced.

3.Nasals: /n, m, ng /. The vocal tract is closed but the velum opens a route to the nasal cavity. The generated voiced sound is affected by both vocal and nasal tract.

4.Tremulants: /r/. Top of the tongue is vibrating quickly (20-25 Hz) against the alveoral ridge causing voiced sound with an effect like amplitude modulation.

5.Laterals: /l/. The top of the tongue closes the vocal tract leaving a side route for the air flow.

6.Semivowels: /j, v/. Semivowels are almost like vowels, but they are more unstable than and not as context-free as normal vowels.

The consonant categories are summarized in Figure 1.8. For example, for phoneme /p/, the categorization will be unvoiced bilabial-plosive.

Consonants	labial		dental alveoral			palatal	velar	laryng.
	bi-lab.	labio-dent.	pro	medio	post			
plosive (tenuis)	p		t				k	
plosive (media)	b			d			g	
fricative (sibilants)			s					
fricative (spirants)		f						h
nasal	m		n				ŋ	
tremulant			r					
lateral			l					
semivowel		v				j		

Figure 1.8 Classification of Finnish consonants.

When synthesizing consonants, better results may be achieved by synthesizing these six consonant groups with separate methods because of different acoustic characteristics. Especially the tremulant /r/ needs a special attention.

IX.Implementation of Text to Speech Synthesis:

In text to speech module text recognised by OCR system will be the inputs of speech synthesis system which is to be converted into speech in .wav file format and creates a wave file named output .wav, which can be listen by using wave file player.

Two steps involved in text to speech synthesis

1.Text to speech conversion

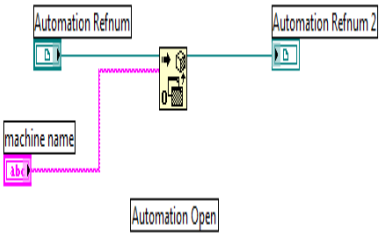
2.Play speech in .wave file formate

1.Text to Speech Conversion:

In the text speech conversion input text is converted speech (in LabVIEW) by using automation open, invoke node and property node will be described below.

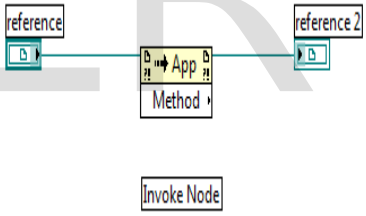
a.Automation Open:

Returns an automation refnum, which points to a specific ActiveX object. In Text to Speech VI, it gives refnum for Microsoft speech object library.



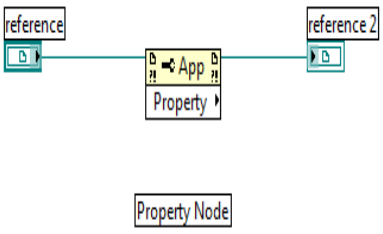
b.Invoke Node:

Invokes a method or action on a reference. Most methods have associated parameters. If the node is configured for VI Server Application class or Virtual Instrument class and reference is unwired, reference defaults to the current Application or VI.



c.Property Node:

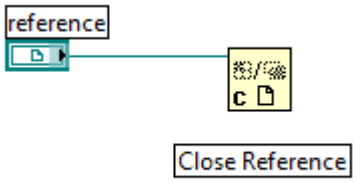
Gets (reads) and/or sets (writes) properties of a reference. The Property Node automatically adapts to the class of the object that you reference. LabVIEW includes Property Nodes preconfigured to access VISA properties and ActiveX properties.



d.Close Reference:

Closes a refnum associated with an

open VI, VI object, an open instance of LabVIEW, or an ActiveX or .NET object.



Some sub vis are used in text to speech conversion like wise rate volume vi and status vi are described below

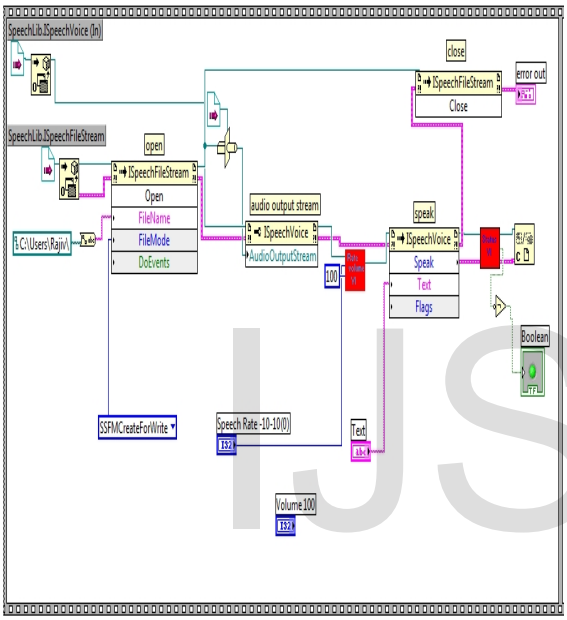


Figure 1.9 Block diagram text to speech Synthesis.

2.Play Speech in Wave File Player

This VI illustrates how to playback a wave file using sound board, information such as file name, sound quality rate & bits/sample are also displayed. We are using this VI to listen the speech generated by text to speech conversion.Various Functions used in wave file player.vi.

X.CONCLUSION

The Optical Character Recognition deals with recognition of optically processed characters. Reliably interpreting text from real-world photos is a challenging problem due to variations in environmental factors even it becomes easier using the best open source OCR engine.

XI.FUTURE SCOPE

Our next works with OCR Mobile Application will include the improvement of the results by the use of table boundaries detection techniques and the use of text post-processing Techniques to detect the noise and to correct bad-recognized words. OCR application will also display the signatures and the other symbols as it is in the document. It will also update its features including the translation of one language to another. So that it will helpful for people from other countries who can't understand the local language.

References:

1.T. Dutoit, "High quality text-to-speech synthesis: a comparison of four candidate algorithms," Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on, vol.i, no., pp.I/565-I/568 vol.1, 19-22 Apr 1994.

2.B.M. Sagar, Shobha G, R. P. Kumar, "OCR for printed Kannada text to machine editable format using database approach" WSEAS Transactions on Computers Volume 7, Pages 766-769, 6 June 2008.

3.G. Nagy, "At the frontiers of OCR," Proceedings of the IEEE, vol.80, no.7, pp.1093-1100, Jul 1992.

4.Landt, Jerry. "Shrouds of Time: The history of RFID," AIM, Inc., 31 May 2006.

5.R.C. Palmer, "The Bar Code Book," Helmers Publishing.

6.Mandell, Lewis. "Diffusion of EFTS among National Banks: Note", Journal of Money, Credit and Banking Vol. 9, No. 2, May, 1977.

7.Bergeron, P. Bryan (1998, August). "Optical mark recognition," Postgraduate Medicine online. June 7, 2006.

8.I. Witten, "Principles of Computer Speech.", Academic Press Inc., 1982.

9.K. Kleijn, K. Paliwal (Editors). "Speech Coding and Synthesis," Elsevier Science B.V., The Netherlands, 1998.

10.Y. Sagisaga, "Speech Synthesis from Text," 1998.