Jurnal Elektronik Ilmu Komputer Udayana
Volume 8, No 3. February 2020

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Low Filtering Method for Noise Reduction at Text to Speech Application

Putu Asri Sri Sutanti[a1], Gst. Ayu Vida Mastrika Giri[a2]

[a]Informatics Department, Udayana University
Bali, Indonesia
[1]srisutanti25@gmail.com
[2]vida.mastrika@cs.unud.ac.id

Abstract

Technological developments have greatly encouraged various researchers to develop several studies in the IT field. One branch of research in the IT field is sound synthesis. Some text-to-speech applications, are usually quite difficult to form and are less flexible in replacing existing types of sound. In addition, sometimes accent or how someone speaks is not well represented, so it is quite difficult to form a text-to-speech application by using the desired sound like user voice or other sounds. From the above problems, this research propose an application that can change text into sound or text-to-speech which is more flexible and in accordance with the wishes of the user. From the results of testing that has been done, this system has an accuracy of 70%.

Keywords : *Text to Speech, Sound, Synthesis ,Low Pass,Noise*

1. Introduction

The development of technology today is very rapid. This has encouraged various researchers to develop some research in the field of IT, which is beneficial for the community. According to the digital marketing research institute Emarketer, in 2018, the number of active smartphone users in Indonesia is expected to reach more than 100 million people [3].

One branch of research in the IT field is sound synthesis. Based on the Big Indonesian Dictionary, synthesis is a mixture of various meanings or things so that it is a harmonious unity, while sound is something that is heard (heard) or captured by the ear [7]. Sounds have different characters from each other. Sound characteristics can be expressed by the parameters of the frequency of the basic tone (prominent frequency) and the color of the sound (timbre). The fundamental tone frequency of a sound is identified by the wavelength of sound that propagates in a fixed medium and has the same propagation rate.

Sound synthesis is a process to produce sound (digital sound) by calculating the value of each sound sample [10]. Sound synthesis is a developing field and is in great demand for research. As for some examples of sound synthesis implementations such as sound generation of traditional and modern music instruments, text to speech, voice changers and others.

Formation of sound from the text feature or commonly called text-to-speech can be done using voice samples, where later the existing sound samples are used to reconstruct each sound in accordance with the existing words.

Text-to-speech (TTS) application is one of the technologies that can meet human needs in the information field. The development of this technology is to menaturikan modeling of human language more precisely. So that various information needs can be met easily [12].

Some text-to-speech applications are usually quite difficult to form and are less flexible in changing existing types of sounds. In addition, sometimes the accent or way of speaking is not well represented, so it is quite difficult to form a text-to-speech application using the desired sound like the user's voice or other voices.

In the study " Rancang Bangun Aplikasi Text to Speech Bahasa Indonesia " written by Sudibyo P. Arbie, making text to speech applications by detecting words as they are in the database requires several sound samples to get the expected results. In addition, to get maximum sound results, the recording is done in a quiet room and without sound reflection [14].

In the study " Pengembangan Aplikasi Text-to-Speech Bahasa Indonesia Menggunakan Metode Finite State Automata Berbasis Android " written by Rieke Adriati W, an Indonesian text-to-speech application that utilizes Google's API for TTS was proposed. Existing words are beheaded according to syllable patterns determined by the Finite State Automata (FSA) method, where English data are used as results for Indonesian by searching for words that match their pronunciation [13].

Research conducted by Achmad Jaka Dwena Putra in the "EXREAD, Aplikasi Pembaca Naskah Ujian Bagi Tuna Netra" produced the conclusion that the application of text-to-speech turned out to have greater benefits, one of which was for the Blind, where text-to-speech - speech is able to become a new alternative to Braille letters for the Blind [1].

Christian's research, discusses how a text-to-speech system can convert text from a word into audio which also contains the pronunciation text and audio context. This conversion can be done by analyzing the previous word. The system produces audio results that can be played according to the words the user wants. However, this research is lacking, namely the sound produced by the system is still unclear in the pronunciation of words, such as the pronunciation of the letters "k", "b" and "d", besides that there are still a few words that cannot be pronounced by the system produced by the data the sound used is still lacking.

From the problems above, the author suggests a system that can change text into voice or text-to-speech that is more flexible and in accordance with the wishes of the user. This application will split each letter sound in a word, and will combine existing sound data to form the sound of a word. The limitation in this paper is that the input of text can only be one word to two words and use English.

2. Research Method

In this research, I am using two datasets in the form of sound data and data word dictionary. Where is the data beep, I get from my recorded voice which contains chunks of letters. Whereas, for the word dictionary data used for decapitating letters in each word according to pronunciation, it is obtained from the CMU dictionary. The Carnegie Mellon University (CMU) Pronouncing Dictionary is an open source machine-readable dictionary for North American English that contains more than 134,000 words and pronunciations [15].

Low-pass filter is a filter that only passes frequencies lower than the cut-off frequency (fc). Above that frequency the signal will be muted [12]. Low-pass filters pass low frequency signals but attenuate signals with frequencies higher than the cutoff frequency [11].

In Figure 1. below, a passband refers to the frequency that is passed, while a stopband refers to the frequency being held and the transition band located between the two.

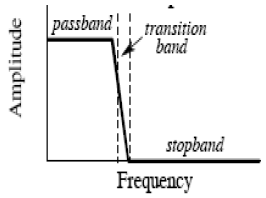

Figure 1. Low Pass Filter

Text to Speech synthesis system consists of 3 parts, namely text pre-processing, prosody generation and concatenation [17].

p-ISSN: 2301-5373
e-ISSN: 2654-5101

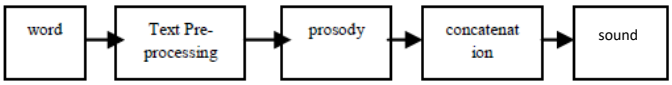

Figure 2. Diagram of a text system to speech synthesis

In the Text Pre-processing section, there is a conversion of input in the form of text into a diphone (a combination of two phonemes). The process of generating prosody is very concerned about the character of human speech signals, to get a more natural speech. Prosody is the change in pitch value (basic frequency) during the sentence pronunciation or pitch as a function of time [17]. The following are the stages of research that will be carried out:

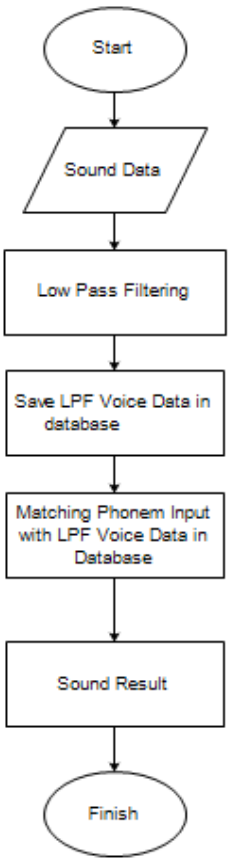

Figure 3. Flowchart Research Stages

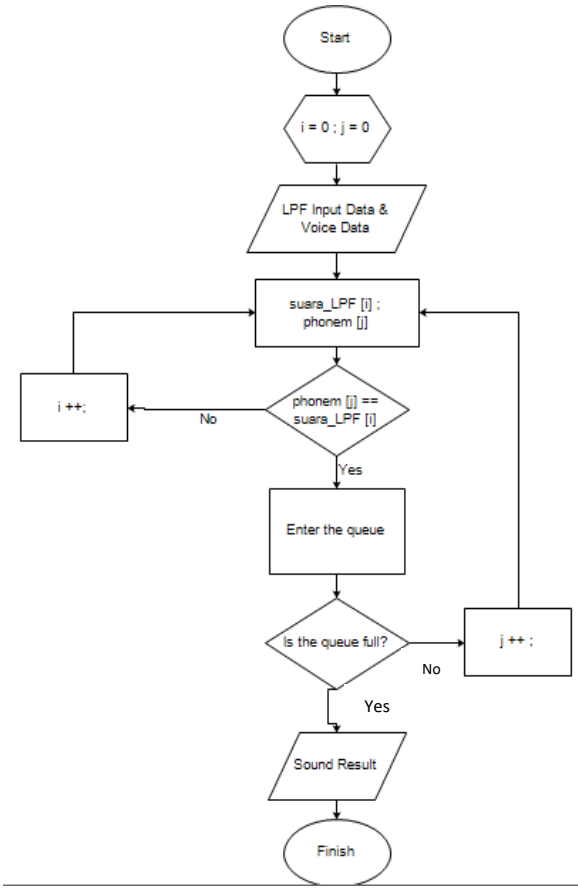

Figure 4. Flowchart Matching Phonem input with LPF Voice Data

2.1. Dataset

The data set that I used in this research is a sound or voice data set that I recorded myself and a data dictionary set of words from the CMU dictionary containing 134,000 words. Where in this sound dataset, it will be stored with the same name as the pronunciation of the word. For example, the word "buy" which has the pronunciation of "[b], [ay]". Then, the sound will be saved with the names b.wav and ay.wav. Because the sound recorded has noise, then to reduce the noise using a low-pass filter.

Table 1. Some Sound Data and CMU Dictionary Data

No.	Sound Data	CMU Dictionary Data
1.	aa.wav	AA
2.	ae.wav	AE

3.	ah.wav	AH
4.	ao.wav	AO
5.	aw.wav	AW
6.	ay.wav	AY
7.	b.wav	B
8.	ch.wav	CH
9.	d.wav	D
10.	er.wav	ER
11.	ey.wav	EY
12.	g.wav	G
13.	jh.wav	JH
14.	l.wav	L
15.	m.wav	M
16.	ow.wav	OW
17.	s.wav	S
18.	t.wav	T
19.	uh.wav	UH
20.	y.wav	Y
21.	z.wav	Z
22.	…	…

2.2. Low Pass Filter

At this stage, a low-pass filter is used to reduce noise from the audio recorded by each letter in the word. Where the frequency used as a limit to reduce noise by 500 Hz. Thus, the sound frequency below 500 Hz will be left while the sound frequency above 500 Hz will be muted. After reducing noise from each dataset, it will be decapitated for each letter sound in the input word. Then the sound of each letter will be merged so that the pronunciation of a word will be formed.

2.3. Merge sounds for each letter

At this stage, the sound file that is saved with the same name as the result of decapitation, will be called and joined in the program. So that sounds are formed for one word. To combine the sounds of each letter using the diphone concatenation method by combining previously recorded sound segments. Each segment is a diphone (a combination of two phonemes). Formation of speech in synthesizing speech using diphone concatenation method in principle is done by arranging a number of corresponding diphone so that the desired speech is obtained.

Before, to the sound merging stage to form a one-word sound. There are several stages, the program will read the data dictionary set of words contained in the CMU dictionary. After that, the input word from the user will be matched with the data contained in the CMU dictionary. If the input with the data matches, then the word fragment from the input will be adjusted to the sound data that has gone through the filtering stages.

2.4. Accuracy Method

After testing, in this study calculates the level of accuracy of the system that has been made.

$$P(N) = \frac{N(A)}{N(S)} \times 100\%$$

Where :

P (N) = level of accuracy

N (A) = the correct amount of data at the time of testing

N (S) = the sum of all test data

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Second, by using a questionnaire distributed to students of the Informatics Engineering Study Program, Faculty of Mathematics and Natural Sciences, Udayana University. In the questionnaire, contains rating ratings regarding the system.

3. Result

3.1 Filtering

In this filtering process, I use the Low Pass process. Where, this low pass process will muffle sounds that have frequencies above the frequency limit of the low pass, while sounds or sounds that have frequencies above the low pass frequency range. This stage produces clearer sound, because this low pass dampens the noise found in the results of previous sound recordings. Where the sound spectrum will be taken, after that from the spectrum will then be carried out the process of low pass filtering. Here are some spectrum images of 21 sound data that can be seen in Figures 5 through 12.

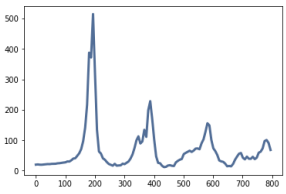

Figure 5. Sound Spectrum from aa.wav

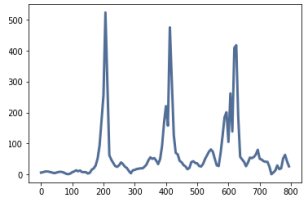

Figure 6. Sound Spectrum from ae.wav

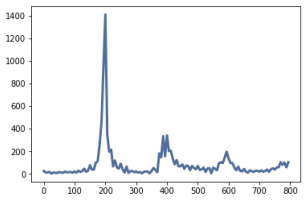

Figure 7. Sound Spectrum from ah.wav

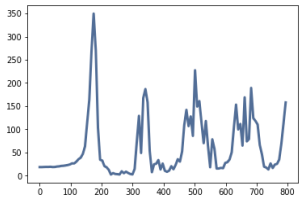

Figure 8. Sound Spectrum from aa,wav

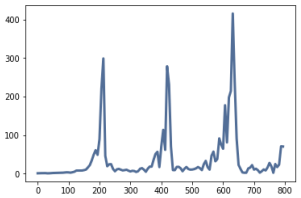

Figure 9. Sound Spectrum from b.wav

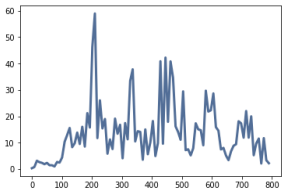

Figure 10. Sound Spectrum from ch.wav

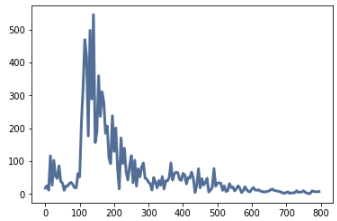

Figure 11. Sound Spectrum from d.wav

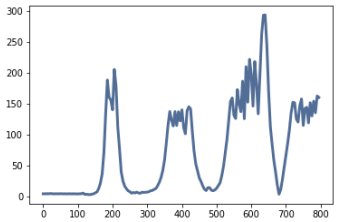

Figure 12. Sound Spectrum from er.wav

The pictures above are some spectrum images of the sound waves used. Where, each sound has a wave of 800 Hz. After that, the next step is to do filtering that produces smoother sound results compared to before. The results of sound filtering will be saved in .wav format in the sound folder so that it can be connected to the program.

3.2. End Result

In this research, I can only convert one word into a sound that can say the words entered by the user. As for some of the words that I have tested in this program:

Table 2. Research result

No.	Word	Pronunciation	Result	
			Success	Failed

1.	buy	[b][ay]	√	
2.	tie	[t][ay]	√	
3.	Hi	[hh][ay]	√	
4.	Hello	[hh][ah][l][ow]	√	
5.	No	[n][ow]	√	
6.	Now	[n][aw]	√	
7.	Allert	[ae][l][er][t]	√	
8.	Allergy	[ae][l][er][jh][iy]	√	
9.	Alles	[ey][l][z]	√	
10.	Alleva	[aa][l][ey][v][ah]	√	
11.	Nice	[n][ay][s]	√	
12.	so	[s][ow]	√	
13.	My Book	[m][ay][b][uh][k]	√	
14.	Good	[g][uh][d]	√	
15.	God	[g][aa][d]	√	
16.	Much	[m][ah][ch]	√	
17.	Bring	[b][r][ih][ng]	√	
18.	Cook	[k][uh][k]	√	
19.	Mom	[m][aa][m]	√	
20.	My door	[m][ay][d][ao][r]	√	

From the table above, it can be seen that in this study it has been able to develop a text-to-speech system from previous research. Where, the system can read a maximum of two words in English and testing is done as much as 20 times.

From the results of tests that have been done, this system has an accuracy of 70%. In addition, I conducted a questionnaire to several students at Udayana University, where to find out the response given about this system, by giving a rating, namely:

1. Not good
2. Not Good
3. Good Enough
4. Good
5. Very Good

Following are the results of the questionnaire that was conducted:

Table 3. Level of Satisfaction with the System

No.	Name	Rating					
		1	2	3	4	5	
1.	I Made Tangkas K.Y				√		
2.	Putu Rikky M.P					√	
3.	Yoel Samosir					√	
4.	I Putu Harta Yoga				√		
5.	Putu Indah Pradnyawati				√		

p-ISSN: 2301-5373
e-ISSN: 2654-5101

6.	Frisca Olivia Gorianto			√		
7.	Kiki Prebiana				√	
8.	Maria Okta Safira			√		
9.	Devin Reness Noak			√		
10.	Sofia Shieldy Budhiono			√		

From the results of the questionnaire, the 10 students obtained a level of satisfaction for this system which is 7 points for a pretty good rating and 3 points for a good rating.

4. Conclusion

From this study, the system was able to convert text to sound, with an accuracy rate of 70%. This research has been able to develop a tet-to-speech system from previous research and the system has been able to convert text to sound with two English words.
It is hoped that for further research, it can further develop existing systems and use suitable audio files.

Reference

[1] Achmad Jaka Dwena Putra.2018. EXREAD. Aplikasi Pembaca Naskah Ujian Bagi Tuna Netra. STMIK Bumigora Mataram. Prosiding PKM-CSR , Vol. 1.
[2] Agus Kurniawan. Reduksi Noise Pada Sinyal Suara dengan Menggunakan Transformasi Wavelet.
[3] Ahmad Subki.2018. Membandingkan Tingkat Kemiripan Rekaman Voice Changer Menggunakan Analisis Pitch, Formant Dan Spectogram.Yogyakarta. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)
[4] Aris Tjahyanto, 2011. Model Analysis-By-Synthesis Aplikasi Pembangkit Suara Gamelan Sintetik.
[5] Christopher Brian Fleizach.2015.Intelligent Text-To-Speech Conversion.US.United States Patent.
[6] F. Ulaby, Fundamental of Appied Electromagnetics, USA: Prentice Hall.
[7] Kamus Besar Bahasa Indonesia (KBBI).Diperoleh pada 13 Mei 2019, dari https://kbbi.web.id/
[8] M. Ichwan.2018. Implementasi Metoda Unit Selection Synthesizer Dalam Pembuatan Speech Synthesizer Suara Suling Recorder. Institut Teknologi Nasional Bandung. MIND Vol.3
[9] Nanik Suciati. Perangkat Lunak Untuk Sintesis Suara Gerakan Benda Padat. Institut Teknologi Sepuluh Nopember.
[10] Neilcy T. Mooniarsih. Desain dan Simulasi Filter FIR Menggunakan Metode Windowing.
[11] P. Hongmei, 2009. Optimization design of UWB passive bandpass filter's standing wave ratio, in IEEE.
[12] Rieke Adriati W. 2016. Pengembangan Aplikasi Text-to-Speech Bahasa Indonesia Menggunakan Metode Finite State Automata Berbasis Android. JNTETI, Vol. 5.
[13] Sudibyo P. Arbie. 2013. Rancang Bangun Aplikasi Text to Speech Bahasa Indonesia. Manado. e-journal Teknik Elektro dan Komputer
[14] The CMU Pronouncing Dictionary.Diperoleh 14 Mei 2019 dari http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[15] Yining Chen, Voice Conversion with Smoothed GMM and MAP Adaptation.
[16] Zonda Rugmiaga. Pembangkitan Prosody Pada Text To Speech Synthesis System Untuk Penutur Berbahasa Indonesia. Institut Teknologi Sepuluh Nopember (ITS) Surabaya.