Journal of Engineering and Applied Sciences 15 (3): 742-751, 2020
ISSN: 1816-949X

A Review Paper Implementation of Indonesian Text-to-Speech using Java

Tito Waluyo Purboyo and Rifki Wijaya
Department of Computer Engineering, Faculty of Electrical Engineering,
Telkom University, Bandung, Indonesia

Abstract: Text-to-speech represent the convert method from text to voice. With this method enable the computer to alter a sentence in one language become the voice form. This technological able to assist human of accomplishment of information requirement instantly. Through this technological aid in one activity session, human able to get the information at one blow conduct the other activity without having to focused at which is being read. In this study, audio data got by the recording process and yield the phoneme data that it is kept in format of Windows PCM (.wav) with the quantization equal to 65.536 level quantization. Research will be done by using Java programming language as assistive appliance. Method of intake voice data adapted by a common method of dismemberment of vowel and consonant in Indonesian. End result from this final duty is an application of text-to-speech Indonesian base on the Java. Voice quality yielded later, then analyzed through MOS method.

Key words: Text-to-speech, Java, Indonesian phoneme, programming language, intake voice, MOS method

INTRODUCTION

Speech is media of communication important. Speech synthesis coveted human ago. Effort to achieve target started around hundreds ago, started from process mechanic until electric synthesis. Text to speech in Indonesia the first time developed, since, 2000 years. Various method applied this application but nature of speech still obstacles. Research about fake speech still walking slowly, based about it research of technology Text To Speech (TTS) did by researcher as contribution form on knowledge.

Automatic conversion of text into speech is text to speech synthesis that resemble to the maximum extent possible, native speakers of the language that reads the text. Text-to-speech synthesizer is a way of computers to talk to you. First, text to speech system gets string text as an input and then an algorithm called text to speech engine analyzes text. The algorithm then processes text and synthesizes speech with mathematical models. Text to speech engine usually produces data sound in audio format.

Converting an arbitrary text into an appropriate waveform is the purpose of the text to speech system. There are two main components of the text system to speech, it is text processing and speech-making. The purpose of the text processing component is to process the input text provided and generate the appropriate sequence of phonemic units. This phonemic unit is manifested by a sound generation component, either by synthesis of parameters or unit selection of large corpus. For speech synthesis that sounds natural, it is essential that the text processing component produces a phoneme sequence units match to arbitrary input text.

There are two main phases in text to speech synthesis procedure. First, text analysis, it is when text input is transcribed into phonetic or other linguistic representations. Second, generation of speech waveforms, in which output is generated from this prosodic information and phonetic. These two phases of text to speech synthesis procedure are usually called high and low synthesis. Input text may be for example data from a word processor, ASCII standard from e-mail, scanned text from a newspaper or mobile text messaging. String character is analyzed into a phonetic representation. This phonetic representation is usually a series of phonemes with additional information (correct intonation, pressure and duration). The speech sound is generated with a low level synthesizer with information from a high level synthesizer. The artificial sounds production like speeches has a long history. It is documented mechanical efforts dating to the eighteenth century.

It's hard to convince an end user that the input to the text to speech system is not a phonemic sequence but a raw text as it is available on documents, blogs, news sites etc. That contain the required text like native scripts, font encodings and non-standard words like address, numbers, currency etc. Most of the related issues in building text to speech for new languages are tied to real-world text handling. The current text to speech system in English

Corresponding Author: Tito Waluyo Purboyo, Department of Computer Engineering, Faculty of Electrical Engineering, Telkom University, Bandung, Indonesia

and well-researched languages uses a rich set of linguistic resources such as word morphological analyzer, disambiguation, letter-appropriate rules, part of speech tags, stress patterns, syllabification in one form or one another to build text processing components from the text to speech system. But for minority languages (which are not well studied lacks sufficient linguistic resources) this involves some complexity ranging from the accumulation of corporate texts in a digital and process able format. The linguistic component is not available in a rich way for all languages of the world. In the practical world, minority languages including some Indian languages lack of the linguistic components.

Research wants to use Java programming tool as interface. Audio data graph is recorded with Cool Edit Pro 2.0 tool and save in Windows PCM mono format with 16 bit quantity of wave extension or equivalent to 65.536 quantization level. This Java programming format can read audio data. The purpose of this study is to make text to speech system in Indonesia language to make interface for desktop user. The benefit of this study is application of text to speech use input text typed directly on interface of desktop computer. This application made for management text to speech, nothing reconfiguration hardware on computer, audio data will be used have format Windows PCM mono with extension wave form 16 bits quantization or equivalent with 65.536 quantization level.

MATERIALS AND METHODS

Sound resulting from tool of speech human: In forming sound language 3 factors such as energy source, tool of speech caused vibrate and change vibration cavity. Airflow from the lungs can open both vocal cords that close together produces certain sound features (Fig. 1). Movement opens and closing the vocal cords caused the air around the vocal cords vibrate. At the time the air from the lungs is exhaled, the two bands the sound can be docked or stretched.

Representation of speech signal: Speech signals are signals that change over time with relatively slow change speeds. At short intervals (between 5 and 100 msec), it has fixed characteristics. At longer intervals it has characteristics that vary according to the phrase being spoken. Figure 2 shows a signal snippet for 100 mse, so that, the entire image shows a 500 min long speech signal.

A fairly general framework based on a pattern recognition approach to voiced-unvoiced-silence classification has been described in which a set of measurements are made on the interval being classified and a minimum non-euclidean distance measure is used to select the appropriate class (Al-Hashemy and Taha, 1988). The parts or components of the speech signal are classified into three different states, namely; Silence (S), the state at which no speech is spoken; Unvoiced (U), the state at which the vocal cord does not vibrate, so, the resulting sound is not periodic or random; Voiced (V), the state at which the vibrations occur in the vocal cord, resulting in a quasi-periodic sound.

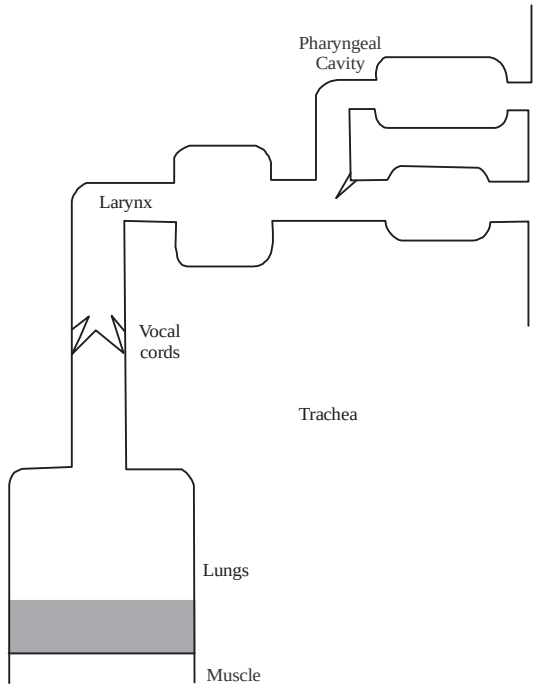

Fig. 1: Model of production speech human system

Table1: Alphabet pronunciation in Indonesia

Alphabet		Name	Alphabet		Name	Alphabet		Name
A	a	a	J	j	je	S	s	es
B	b	be	K	k	ka	T	t	te
C	c	ce	L	l	el	U	u	u
D	d	de	M	m	em	V	v	fe
E	e	e	N	n	en	W	w	We
F	f	ef	O	o	o	X	x	eks
G	g	ge	P	p	pe	Y	y	ye
H	h	ha	Q	q	ki	Z	z	zet
I	i	i	R	r	er			

Language system

Alphabet: Alphabets are used in the spelling of Indonesia language consists the next alphabet. Name each alphabet included beside it (Table 1).

Vocal: Vocal is the sound of language whose air currents are not experience obstacles and their quality is determined by three factors: high low tongue position, increased tongue portion and the shape of the lips on the vocal formation. In Indonesian there are six vowels: /i/,/

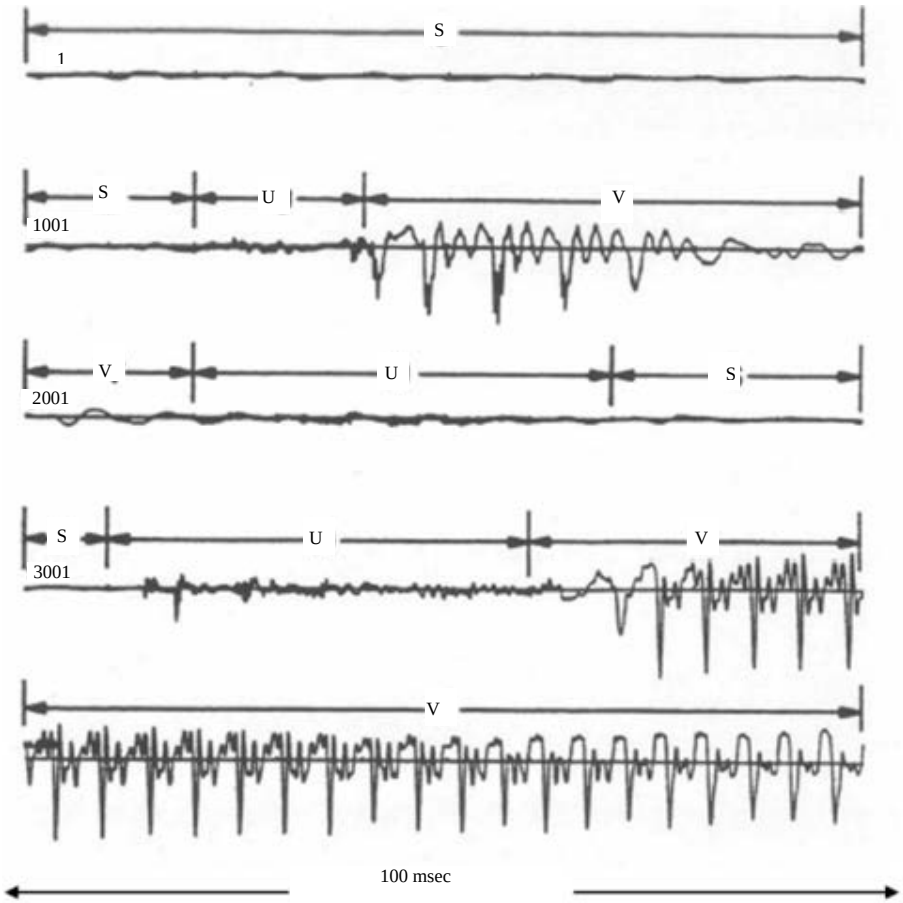

Fig. 2: Speech signal

e/, //, /a /, / u/ and /o/. The six Indonesian vowels can occupy the position beginning, middle or end of syllable.

Consonant: Sound of consonant made different ways. On consonant pronunciation, 3 factors included: condition vocal cords, touch or approach various speech tools and this step of speech tools touch and approach (Alwi, 2007). Consonant can categorized as voiced consonant and unvoiced consonant. According to articulation, consonant in Indonesia language can categorized by 3 factors condition vocal cord, articulation areas, articulation ways. Based condition vocal cords, consonant can voiced or unvoiced. Alphabet symbolize consonant in Indonesia language consists of alphabets b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, y and z.

Combined alphabet-consonant: In Indonesia language be found 4 combined alphabet that are kh, ng, ny and sy. Each one of them symbolize one consonant sound.

Phoneme allophone and grapheme: Phoneme is language minimal distinguish form and meaning of words. Allophone is phoneme do not distinguish meaning of words. Grapheme is representation phoneme in alphabet form.

Diphthong: Diphthong is a vocal that changes its quality at the time of pronunciation. In Indonesian there are three pieces diphthongs, i.e.,/ay/,/aw/and/oy/ which can be written, respectively: ai, au and oi. The three diphthongs are phonemic in Indonesian. Both vowels on the diphthong symbolize one inseparable vowel sound. The sounds of au and ai in the word leaf and main for example are not diphthongs because either a or u or i each get the same (almost) equal pressure and form a separate syllable, so that, the word leaf and main each consist of two syllables: da-un, ma-in.

Consonant cluster: Consonant clusters are a row of two or more consonants belonging to the same syllable. The [pr] sound of the word practice is a consonant cluster as well as pl on plastics, tr in literature and str on the structure. The separation of sounds in the word is prak-tik, plas tik, sas-tra and struk-tur.

Syllables structure words and consonant cluster: The syllables in Indonesian can consist of V, VK, KV, KVK, KVKK, KVKKK, KKV, KKVK, KKKV, KKKVK, KKVKK. Vocals and consonants fill the syllables pattern on V until KVKKK are generally, any vocals and consonants.

However, for the KKV pattern until KKVKK the scope is more limited. If these two consonants are in the same syllable, the first consonant is limited to the resonant constraints /p, b, t, d, k, g/ and the fricative consonant /f, s/ while the second consonant is limited to the consonant /r/ or / l, w, s, m, n, f, t, k / ie: /pl/, /bl/, / kl /, / gl /, / fl /, / sl /, / pr /, / br /, / tr /, / dr /, / kr /, / gr /, /sr/, /ps/, /sw/, /kw/, /sp/, /sm/ , /sn/, /sk/, /pt/, /ts/, /st/.

If three consonants lined up in one syllable, the first consonant is always /s/, the second /t/, /p/ or /k/ and the third /r/ or /l/. Namely: /str/, /spr/, /skr/, /skl/.

Beheading words: Beheading words on basic words do as the follows: if in the middle of the word there is a sequential vowels, the beheading is done between the two vowels. The diphthong letters ai, au and oi are never divorced, so that, the beheading is not done between the two letters.

If in the middle of the word there are consonants including consonants of consonants, between two vowels, beheaded before consonants. If in the middle of the word there are two consonant letters in sequence, beheading is done between the two consonants. Combine alphabet-consonant are never separated. If in the middle word there are three alphabet or more consonants, the beheading is done between the first consonant and the second consonant.

Basic components of text to speech system: TTS system will produce speech signal automatically via. grapheme transcript to phoneme on sentences given. There are procedure Text to Speech (TTS) synthesis consists of 2 main phases which are:

- Text analysis, text translated in phonetics or language representation
- Evocation sound wave produced from phonetic

Simply, procedure is displayed on Fig. 3 and 4. The Input text is string then prepared and analyzed to phonetic form which is usually taken from the phoneme with some additional information. For TTS forming needed the shortest part of the voice signal such as syllables, phoneme or the shortest segment.

Digital speech signal: Human speech signals are known as analog acoustic signals. In signal processing, the analog signal is converted into a digital signal through a digitization process. In this final project the method used is PCM (Pulse Code Modulation).

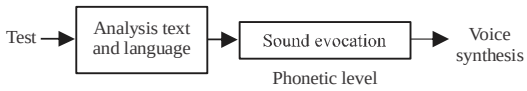

Fig. 3: Procedure TTS simple synthesis (Khalifa *et al.*, 2011)

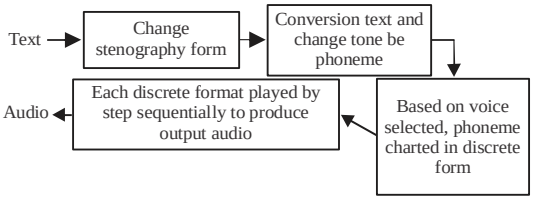

Fig. 4: Step of forming TTS

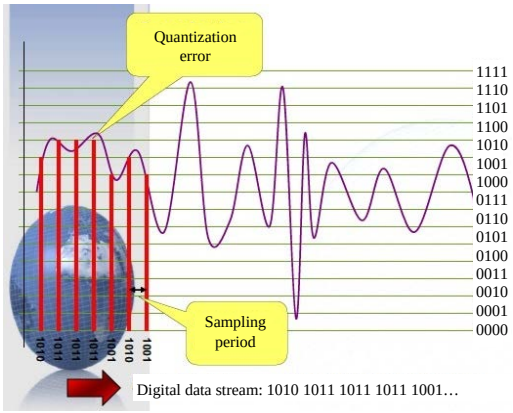

Fig. 5: Sampling process

There are two processes in digitizing analog signals namely sampling and quantization and coding. The voice signal can be converted into electrical signals through the microphone.

Sampling: The minimum value of sampling frequency is called Nyquist rate. If the sampling frequency (fs) is less than the Nyquist rate it will cause distortion aliasing that is damaged component of high frequency signal. For speech signal processing, the usual sampling frequency is 6 until 16 kHz (Fig. 5). In sampling process, the analog signal s (t) is converted into a series {si) = fs (iT)} at time ti = iT where i is the integer number. T is called sampling time while fs = 1/T is the sampling frequency.

An analog signal can be represented by its discrete signal if its sampling frequency is at least twice the highest frequency found in analog signals. Fs = 2 Fm.

Quantization: Quantization is done to make the signal discrete in amplitude. If the signal is quantized by using n bits then the signal amplitude is divided into 2n quantization level. For uniform quantization, the step-size

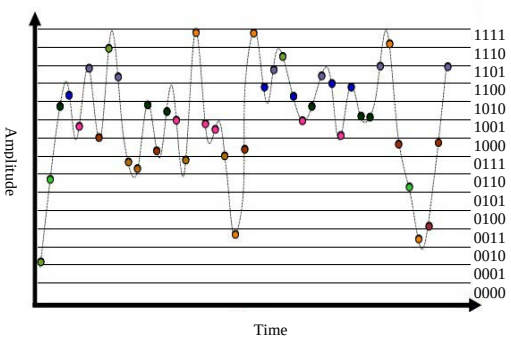

Fig. 6: Quantization

level of quantization is equal there are two things to consider in quantization techniques that minimize data speed and maximize data quality (Fig. 6).

The data rate is measured from the number of bits per second (bits per second), thus, minimizing the data rate means reducing the number of bits used to represent the signal in each sample. Maximizing data quality means generating digital waves that can be converted back to analogue with a small error rate. The quantization process can result partial loss of information called quantization error. Quantization error is the difference of the quantization signal and the original signal expressed as follows:

$$e(nT) = s(nT)\text{-}Sq(nT)$$

Sound signal quality is measured through Signal-to Ratio (SNR) in decibels (dB) that is:

$$SNR = \frac{\sum_{i=1} S^{32}(i)}{\sum_{i=1} e^{2}(i)}$$

Low SNR values indicate low sound quality. Quantization errors can be reduced by minimizing quantization step-size, so that, the difference between the sampling level and the quantization level is not too large, therefore, increasing the data rate as it will increase the number of bits per sample required to represent the signal.

Encoding stage: If at the quantization stage, the value of each sampling is not in a binary number, then at the coding stage each of the quantization levels will be presented in binary form. The trick is to provide the codes at each level of quantization and then the codes are represented in the binary. For more details below described all stages of PCM signal formation.

From the figure it can be seen that to send the analog signal then sent is the value of sampling 1.3, 3.6, 2.3, ... and so on. But if the quantization signal sent is 1.5, 3.5, 2.5, ... and so on. If it will be transmitted PCM signal then sent in the form of a series of pulses, therefore, called PCM or Pulse Code Modulation.

File wave: To be able to store voice or speech in the form of computer data required a form of file storage that can be either wave or MIDI. The wave file format is a subset of Microsoft's RIFF specification for the storage of multimedia files (Wakiyama *et al*., 2010). Before a text to speech can be saved in the form of a file must be changed first speech signal into digital form.

A PC computer for storing or recording a sound requires a multimedia device that acts as an analog (analog) to digital device that converts from analog to digital signal to be stored into a wave/midi file and vice versa from digital to analog to speak files that have been saved. In the wave file storage the voice data signal is stored in the 44th bytes until the last byte of the file while the bytes to 0-43 are the file headers.

Text analysis: The first stage of the TTS system is to change the data input to the appropriate form for the synthesizer. At this stage, all letters outside the alphabet such as numbers, acronyms, abbreviations and must be converted into full greeting formats.

Text analysis is made with tables face-to-face. Additional information on adjacent words or letters should be added. For that required a sufficient database, set of rules and reliability of the system against real time. The text analysis module has several stages of the process as follows:

Pre-processing which organizes the input text, so, that it becomes an adjustable form, so that, numbers, abbreviations, acronyms and idioms can be recognized and then transformed into full text as necessary.

Morphological analysis which gives all possible category of speech unit of each word with spell-based. The combined word, the word absorption and inflection will be described as the basic grapheme unit.

Contextual analysis choose words in context, thus, reducing the number of possible categories of other speech units to the amount that can give the correct part of the greeting with adjacent words.

Analysis of syntax and parser prosody, ensuring the existence of spaces and determining the most appropriate text structure which is closer to the realization of prosody.

The phoneme converter to speech functions to generate speech signals based on the phoneme code generated from the previous process. This sub-system should have a library of each speech unit of a language. In systems using phone concatenation techniques, the system must be supported by a phone database that contains the recording of phoneme speech segments. Speech in a language is formed from a set of sounds that may be different for each language, therefore, each language must be equipped with different database phonemes.

The main stages of converting from text to speech can be expressed by a diagram as seen in Fig. 7. The

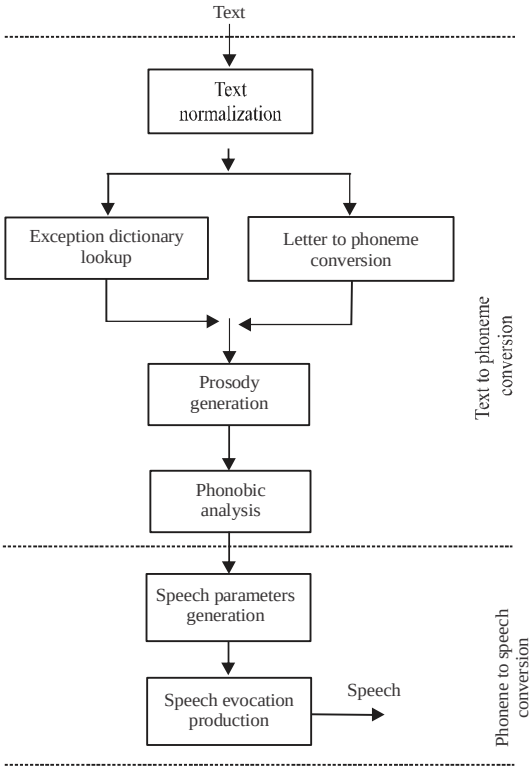

Fig. 7: Sequence conversion text to speech process

normalization stage of the text serves to change all the text of the sentence to be spoken into a text that fully shows how to pronounce it.

The next step is to convert from a text that already fully represents the sentence to be phoneme code. Converting text into phonemes is usually done in two ways. Some of the conversion process can be done with simple conversion rules and is generally, accepted for various conditions. Some other processes are conditional, depending on the letters or neighboring phonemes, there are even translational forms that cannot be found regularity.

Regular conversions can be implemented with a conversion table containing pairs between sequences of letters and phoneme sequences, perhaps even containing only one letter and one phoneme. A more difficult rule is usually implemented with a conversion table that will be applied if the left and right neighbor letters are met. Indonesia includes a clear language of its conversion rules. Most of the words in Indonesian words can be converted into phonemes with clear and simple rules, although, there are conditions that cannot be found in order. For example, the letter symbol e can be pronounced as e ‘pepet’ or e ‘taling’, meaning it must be converted to a different phoneme for different conditions. In the block diagram above conditions that

Table 2: Fragment vocal (Alwi, 1993)

	Word sample		
Alphabet	Early	Middle	End
A/a	Anak	Kantor	Kota
E/e	Ekor	Nenek	Sore
I/i	Ikan	Pintu	Api
O/o	Obat	Kontan	Baso
U/u	Ukir	Tunda	Pintu

Table 3: Consonant fragment (Alwi, 1993)

	Example words		
Alphabet	Early	Middle	End
B/b	Bahasa	Sebut	Adab
C/c	CAKAP	Kaca	
D/d	Dua	Ada	Abad
F/f	Fakir	Kafan	Maaf
G/g	Guna	Tiga	Gudeg
H/h	Hari	Saham	Tuah
J/j	Jalan	Manja	Mikraj
K/k	Kami	Paksa	Politik
L/l	Lekas	Alas	Kesal
M/m	Maka	Kami	Diam
N/n	Nama	Anak	Daun
P/p	Pasang	Apa	Siap
Q/q	Quran	Furqan	Fariq
R/r	Raih	Bara	Putar
S/s	Sampai	Asli	Lemas
T/t	Tali	Mata	Rapat
V/v	Varia	Lava	
W/w	Wanita	Hawa	Waw
X/x	Xenon		
Y/y	Yakin	Payung	
Z/z	Zeni	Lazim	Juz

Table 4: Diphthong

		Example words		
Diphthong	Grapheme	Early	Middle	End
ai	<ay>	Ain	Syaitan	Pandai
au	<aw>	Aula	Saudara	Harimau
oi	<oy>		Boikot	Amboi

can still be handled by the rules are implemented with letter to phoneme conversion block. Irregular conversions are handled by the exception dictionary lookup section.

Table 2 and 3 the result of this stage is a series of phonemes that represent the sound of sentences to be spoken. The prosody generator section will complete each phoneme unit generated with its pronunciation duration and pitch. Duration and pitch data were obtained based on a combination of tables or databases and prosodic models. Symbolically, the results of this section have yielded enough information to produce the desired speech Table 4 and 5.

Language of Java programming: This study use Java programming to implementation text to speech with consideration:

Table 5: Combine of alphabet-consonant

	Example words		
CAC	Early	Middle	End
kh	Khusus	Akhir	Tarikh
ng	Ngilu	Bangun	Senang
ny	Nyata	Hanyut	
sy	Syarat	Isyarat	

- Text based application
- Based windows or GUI based application each platform operation system
- Designing of Java eliminate allocation and deallocation manual
- Java apply true array, eliminate necessities arithmetic pointer dangerous and easy to be wrong
- Eliminating multiple inheritance replaced with interface facilities

To form text-to-speech in Java programming language required five public classes which are:

Engine: This class serves to define an engine's task. An engine must have the ability to normalize, phonemize and retrieve audio data.

Engine impl: This class works to implement the detailed task assignment engine that has been defined in the engine class.

Engine template: Describes the flow or process template of the TTS includes normalization of sentences, tokenization and mapping of sound files. Font tokenizer this class looks for linkages of the n, y, g, k, h, s, y, a, i, o, u in the diphthongs and the consonant letters.

Runner: This class is the program executor.

GUIRunner: This class is a GUI representation.

Atom: This class is for data structures represents the smallest letter. An atom has content and type properties. The example of the letter n is represented by content = n and type = consonant, another example: the letter ng is represented by content = ng and type = consonant, another sample: the letter i is represented by content = i and type = vowel.

Sound file: Data structures that represent sound files. A file has a filename property and its delay.

Currency: This class is a utility to change the numbers to be spelled out.

Font position util: This class serves as a utility to find out the position info of the relative font of the current position.

Letter Indonesia: It is a utility to define the font type whether it is a vowel consonant or any other character.

Player sound file: This class serves as a utility to play a series of pre-prepared sound files.

The description is as follows. GUIRunner runs then the user enters a sentence in the text box. When the user clicks the listen button, GUIRunner creates the engine object, then creates the engine template object. The engine object and the text that was typed by the user are assigned to the engine template object. After that GUI runner ordered the object engine template to start the process. The engine template object instructs the engine object to perform normalization. After the normalization is complete. The engine template object again instructs the engine object to perform tokenization (in order to do tokenization, engine object assisted by font tokenizer object).

After completion of tokenization, the engine template object instructs the engine object to perform the mapping file (in the mapping file, the engine object is assisted by font position util to determine if the letter is positioned -X, -X-, X- or X) the result is a list of sound files ready run by player. Up here the task engine template and engine is done. List of files was then given to the player to play (the application only gives the name of the file to be run, Java already provides the player, we just give the file name to Java).

Audio database: Audio database will be used this study, selected from speech in included phoneme represent phoneme in Indonesia language.

Single pronunciation: Single pronunciation is the sound of pronunciation letters of the alphabet directly without going through the beheading of a word. The alphabetized letters are pronounced 26 alphabets such as: A/a, B/b, C /c, D/d, E/e, F/f, G/g, H/h, I/i, J/j, K/k, L/l, M/m, N/n, O/o, P/p, Q/q, R/r, S/s, T/t, U/u, V/v, W/w, X/x, Y/y, Z/z. This audio data is the same for capital letters and lowercase letters. The symbol of a space is silence obtained by recording under conditions mute.

Fragment vocal: Fragment of vocal alphabet represented by: the result of this fragment is the phoneme for the initial position of the syllable, the middle of the syllable and the end of the term words. Which is then marked with. For starters the syllable becomes x-, for the middle of the syllable to be - x- for the end of the syllable to be -x.

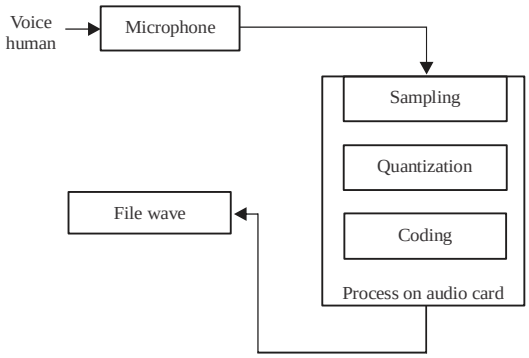

Fig. 8: Process speech digitation

Consonant fragments: In English, consonant fragment are trusted than vowel fragment (Boudelaa, 2015). Consonant letter fragments are represented by:

The result of this fragment is the phoneme for the initial position of the syllable, the middle of the syllable and the syllable end. Which is then marked with "-". For the beginning of the syllable to be x- for the middle of the syllable to be -x- for the syllable end becomes -x.

Diphthong: Diphthong in Indonesian are represented by: the result of this fragment is the phoneme for the initial position of the syllable, the middle of the syllable and the syllable end. Which is then marked with. For the beginning of the syllable to be xx, for the middle of the syllable to be xx-, for the syllable end becomes -x.

Combine of Alphabet-Consonant (CAC): Combine of alphabet-consonant are represented by: the result of this fragment is the phoneme for the initial position of the syllable, the middle of the syllable and the syllable end. Which is then marked with. For the beginning of the syllable to be xx- for the middle of the syllable to -xx- for the end of the syllable to -xx.

Digitation speech signal: Process digitation speech signal described on diagram block (Fig. 8 and 9).

Voice of human: An example of the sound recorded for this final project is the sound of the author with the type of baritone sound. The word to be recorded is chosen according to the phonemic requirement. The phoneme data is taken by cutting method. The process of recording is done in a closed room for avoid any noise. Used 2 way communication through speakers and headset, so that, sound operator can be eliminated.

Microphone: Microphone used is a standard microphone for studio recording process. When the sound signal is converted into an analog signal s (t) with a gain of 20 dB.

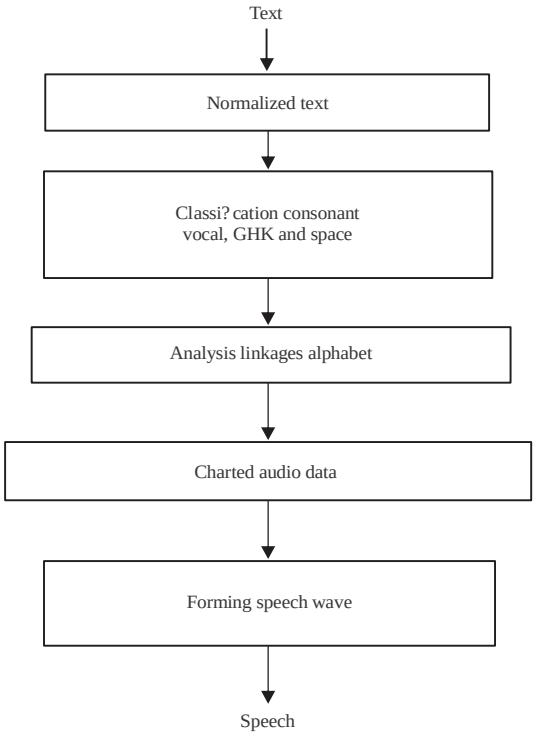

Fig. 9: Text to speech diagram block

The position of the mic is 60°. There is Patti who used 20° (Price *et al*., 1988) for mic but we stick to 60°.

Sampling: In the sampling process, the analog signal s (t) is converted into a series {si) = fs (iT) at time ti = iT, where i is an integer. T is called sampling time while fs = 1/T is the sampling frequency. The window function used is Blackmann-Harris. This function is a standard window function in Cool Edit Pro 2.0 recorder Software:

When M = 1 is a scale factor for window width. To present an analog signal through a discrete signal, the sampling frequency is twice the highest frequency contained in the analog signal.

Fs = 2 Fm for speech signal processing, the sample rate used is 44.100 sample rate. Sampling frequency used ranged from 20-20 kHz.

Quantization: In Cool Edit Pro 2.0 recording Software, quantization of 65,536 quantization levels using 16 bit resolution. The quality produced with this resolution approaches the normal human voice.

Encoding stage: The encoding stage is done automatically by Cool Edit pro 2.0 recorder software

which then formed a series of PCM pulse signal modulation (Pulse Code Modulation). Speech signals can be analyzed with using a spectrograph as shown in Fig. 9. A speech segment that resembles a time domain is more easily differentiated on a spectrograph by looking at the difference in its frequency components.

File wave: Audio files are stored in windows format PCM with wav extension. The selection of this format is tailored to a format that can be read in a Java program. In the wave file storage the voice data signal is stored in the 44th bytes until the last byte of the file while the bytes to 0-43 are the file headers.

Conversion text to speech: Diagram alteration text to be speech on this study described the following as:
Text: Input character generated keyboard and commonly used in the Indonesian language.

Text normalization: In this study, the normalization of characters performed by text normalization. The normalization of the text produces text that is less awkward and more familiar to recipients of the text (Alleva *et al*., 1999).

Classification of vowels consonants GHK and spaces: Knagenhjelm use ANN to classify continuous speech (Knagenhjelm and Brauer, 1990). At this stage the normalized tests are then classified into vowels, consonants, GHK and spaces.

RESULTS AND DISCUSSION

Testing system: Testing performance system can be done through several stages as follows:

- Analysis of word to syllable conversion
- Analysis of wave file incorporation

Covert words to syllables: In the early stages output test system of the conversion program text to syllables will be analyzed according to the rules that are specified. Because in the implementation of the Indonesian text-to-speech program it follows the following syllable slashing rules:

- Diphthong ai, au and oi were never divorced as well as the consonants-kh, ng, ny, sy were never divorced, so the beheading was not done between the two letters
- Each syllable has only one vowel
- If in the middle of the word there are consonants including consonant letters, between two vowels, beheaded before consonants
- If there is a consonant between vowels, then the consonant becomes a syllable of the second vowel
- Nominal normalization becomes a sequential word each starting from units, tens, hundreds, thousands and millions
- For a sentence, between words are given a space break

From the result of the conversion of syllables that have been done then it can be seen that to form syllables that contain many consonants and prefix is difficult to apply because the interconnection between syllables in the use of simpler syllables is different.

Merging file wave of syllables: In the testing phase of the incorporation of this wave film, the results cannot be analyzed directly, so that, the tests at this stage are directly included in the test analysis. From the results of hearings test can be well known how the results of the merger wave file.

Analysis text to speech Indonesian: From the tests that have been done visible that the text to speech is Indonesian based on the relevance letters that have been created can work properly, though the reading has not been intonated. This matter evident from the test results that the average word can be understood by the listener as much as 40.75%, although, there are still some the word is still not understood which is about 19.54%. Text to speech system that has been built is still not perfect, the system has not been able to anticipate the things as follows. The lack of perfect process of beheading words into syllables, so that, not all words can be beheaded according to pronunciation.

The inclusion of a syllable utterance from the user to be stored on the syllabary list is less than perfect, for example, containing a particular intonation wherein the intonation is not appropriate for the word in question. (Eg contains elements of the regional language).

In addition to the terms of the system, the quality of the Indonesian text to speech with syllabic synthesis is influenced by considerable aspects including individual intelligence, educational level, health, the language of the user's home country, the entry or enticement of syllable speech and psychological conditions concerned.

Speech Indonesian with syllabic synthesis is influenced by considerable, intermediate aspects other individual intelligence, educational level, health, the user's home language, entry or entry syllable words and psychological conditions concerned.

CONCLUSION

Based on the tests that have been done against the built system then it can be taken some conclusions is

follows. Indonesian text-to-speech can be built with the incorporation of the correspondence of the letters based on syllable language rules Indonesia is proven with its words generated easily understood by the listener.

Text-to-speech Indonesian merging letter relation has been able to say most of the Indonesian words with easy-to-understand pronunciation meaning he said.

Not all words in the Indonesian language can be converted by this system, especially, words whose beheading does not match the pronunciation or unique way of beheadings (not in accordance with the criteria of the built system). The result of the pronunciation depends on the quality of the existing syllable words. So, the syllable sounds to be stored in the data base should be arranged in such a way that it is expected to have the same loud sound and the same tone.

For the pronunciation of a tribe containing “e” or “e” is still indistinguishable and sometimes still produces ambiguous or inappropriate words.This program can run with a minimum memory requirement of 256 and 1.6 GHz processor.

REFERENCES

Al-Hashemy, B.A.R. and S.M.R. Taha, 1988. Voiced-unvoiced-silence classification of speech signals based on statistical approaches. Appl. Acoust., 25: 169-179.

Alleva, F.A., M.J. Rozak and L.J. Israel, 1999. Text normalization using a context-free grammar. U.S. Patent No. 5,970,449, Patent and Trademark Office, Washington, DC, USA.

Alwi, H., 1993. [General Guidelines for Enhanced Indonesian Spelling]. Grasindo Publisher, Jakarta, Indonesia, (In Indonesian).

Alwi, H., 2007. [Standard Indonesian Language]. Ministry of Education and Culture, Central Jakarta, Indonesia, ISBN:9789794071779, Pages: 475 (In Indonesia).

Boudelaa, S., 2015. The differential time course for consonant and vowel processing in Arabic: Implications for language learning and rehabilitation. Front. Psychol., Vol. 5,

Khalifa, O.O., M.Z. Obaid, A.W. Naji and J.I. Daoud, 2011. A rule-based arabic text-to-speech system based on hybrid synthesis technique. Aust. J. Basic Applied Sci., 5: 342-354.

Knagenhjelm, P. and P. Brauer, 1990. Classification of vowels in continuous speech using MLP and a hybrid net. Speech Commun., 9: 31-34.

Price, P., W.M. Fisher, J. Bernstein and D.S. Pallett, 1988. The DARPA 1000-word resource management database for continuous speech recognition. Proceedings of the International Conference on Acoustics, Speech and Signal Processing ICASSP-88, April 11-14, 1988, IEEE, New York, USA., pp: 651-654.

Sunendar, D. and Sugiyono, 2016. [General Spellings Guidelines for Indonesian]. Kementerian Pendidikan dan Kebudayaan, Central Jakarta, Indonesia, ISBN:978-979-069-262-6, Pages: 78 (In Indonesia).

Wakiyama, M., Y. Hidaka and K. Nozaki, 2010. An audio steganography by a low-bit coding method with wave files. Proceedings of the 2010 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), October 15-17, 2010, IEEE, Darmstadt, Germany, ISBN:978-1-4244-8378-5, pp: 530-533.