International Journal of Software Engineering & Computer Sciences (IJSECS)
ISSN: 2289-8522 ,Volume 1, pp. 85-93, February 2015

DOI: http://dx.doi.org/10.15282/ijsecs.1.2015.7.0007

INDONESIAN TEXT-TO-SPEECH SYSTEM USING DIPHONE CONCATENATIVE SYNTHESIS

Sutarman

[1]Faculty of Information Technology and Business, University Technology of Yogyakarta, Jalan Ringroad Utara, Jombor, Sleman 55285, Yogyakarta, Indonesia
Email: sutarman@uty.ac.id

ABSTRACT

In this paper, we describe the design and develop a database of Indonesian diphone synthesis using speech segment of recorded voice to be converted from text to speech and save it as audio file like WAV or MP3. In designing and develop a database of Indonesian diphone there are several steps to follow; First, developed Diphone database includes: create a list of sample of words consisting of diphones organized by prioritizing looking diphone located in the middle of a word if not at the beginning or end; recording the samples of words by segmentation. ;create diphones made with a tool Diphone Studio 1.3. Second, develop system using Microsoft Visual Delphi 6.0, includes: the conversion system from the input of numbers, acronyms, words, and sentences into representations diphone. There are two kinds of conversion (process) alleged in analyzing the Indonesian text-to-speech system. One is to convert the text to be sounded to phonem and two, to convert the phonem to speech. Method used in this research is called Diphone Concatenative synthesis, in which recorded sound segments are collected. Every segment consists of a diphone (2 phonems). This synthesizer may produce voice with high level of naturalness. The Indonesian Text to Speech system can differentiate special phonemes like in 'Beda' and 'Bedak' but sample of other spesific words is necessary to put into the system. This Indonesia TTS system can handle texts with abbreviation, there is the facility to add such words.

***Keywords*:** diphone; text to speech; concatenative synthesis.

INTRODUCTION

Text-To-Speech (TTS) TTS is a system that converts any text sentence in a specific language to be a speech in the same language. It helps, for instance, to automatize talks such as in telecommunication and multimedia, help the mutes or develop language education (Arman A. A, 2002).

There are two main modules in the TTS synthesizer, namely Natural Language Processing (NLP) and Digital Signal Processing (DSP), (Nur Aziza Azis, Hikmah et al. 2011). Natural language processing module is responsible for conversion of text input into phonetic transcription and prosody information. Prosody information, which includes melody (intonation) and rhythm, is necessary to make the resulting speech sounds natural (not flat/robot-like). The DSP module then transforms the resulting phonetic transcription and prosody information into corresponding speech.

Most of the existing TTS system can be majorly classified as formant synthesizer and concatenative synthesizer. Concatenative synthesizer has become popular in recent years due to its ability to produce natural-sounds output. A slight drawback of this concatenative approach is the need of powerful computational method and sufficient storage. As memory cost has dropped,

it is possible to increase the size of speech database size in order to increase the quality of the speech, as well as improving computational algorithm used in the system. Concatenative synthesizer is divided into two principal approach; diphone concatenation and syllable concatenation (Richard M. and Aulia A, 2013).

Research objects with TTS Indonesian such as: A TTS Indonesian research has been done by (Arman A.A, 2002). It's used MBROLA technology and diphone database making process in Belgium, which created MBROLA technology. An SMS application with Text To Speech of Indonesian on Symbian Operating System for the Blind by (Rommel, E, 2005). Indonesian Text-To-Speech System Using Syllable Concatenation: Speech Optimization by (Richard M. and Aulia A, 2013). Handi D. R. B. and Miftahul H (2011) Text Pre-Processing of Text To Speech Synthesis System for Speak of Indonesian Language.

INDONESIAN LANGUAGE

The Indonesian language, so-called Bahasa Indonesia, is a unity language formed from hundreds of languages spoken in the Indonesian archipelago. It was coined by Indonesian nationalists in 1928 and became a symbol of national identity during the struggle for independence in 1945. Compared to other languages, which have a high density of native speakers, Indonesian is spoken as a mother tongue by only 7% of the population, and more than 195 million people speak it as a second language with varying degrees of proficiency. Approximately, there are 300 ethnic groups living in 17,508 islands, speaking 365 native languages or no less than 669 dialects (Johannes Tan, 2009). At home, people speak their own language, such as Javanese, Sundanese or Balinese, though almost everybody has a good understanding of Indonesian as they learn it in school.

Although the Indonesian language is infused with highly distinctive accents from different ethnic languages, there are many similarities in patterns across the archipelago. Modern Indonesian is derived from the literary of the Malay dialect, which was the lingua franca of Southeast Asia. Thus, it is closely related to Malay spoken in Malaysia, Singapore, Brunei, and some other areas. Concerning the number of speakers, today Malay-Indonesian ranks around sixth in size among the world's languages. The only difference is that Indonesia (which was a Dutch colony) adopted the Van Ophuysen orthography in 1901, while Malaysia (which was a British colony) adopted the Wilkinson orthography in 1904. In 1972, the governments of Indonesia and Malaysia agreed to standardize the "improved" spelling, which is now, in effect, on both sides. Even so, modern Indonesian and modern Malaysian are as different from one another as are Flemish and Dutch (Johannes Tan, 2009).

The standard Indonesian language is continuously being developed and transformed to make it more suitable to the diverse needs of a modernizing society. Many words in the vocabulary reflect the historical influence of various foreign cultures that have passed through the archipelago. It has borrowed heavily from Indian Sanskrit, Chinese, Arabic, Portuguese, Dutch, and English. Although the earliest records in Malay inscriptions are syllable-based written in Arabic script, modern Indonesian is phonetic-based written in Roman script (George Quinn, 2001). It used only 26 letters as in the English/Dutch alphabet.

Indonesian know the writing language and spoken language. sometimes there some differences in the two types of these languages. In spoken language, known term phoneme, which is the smallest unitary language that can differentiate meaning. In written language, phonemes denoted by the letter. In other words, letter is writing of phonemes. The following is Indonesian concept-based guideline's general Indonesian spelling perfected. Alphabet, Alphabet used in Indonesian consists of 52 letters, namely 26 uppercase (AZ) and 26 lowercase (az).

Phoneme, Phoneme is a linguistic term and is the smallest unit in a language that they can show a difference of meaning. The Indonesian has 35 phonemes. Diphone, Diphone is a combination of two phonemes Indonesian. Total in Indonesian diphone of approximately 1024 diphone.

TEXT TO SPEECH SYNTHESIS SYSTEM

Text-to-Specch (TTS) is an automatic production of sound, which transforms grapheme transcriptions into phonemes from sentences into utterance/speech (Dutoit and Leich, 1993). A speech synthesizer or text to Speech, in principle, consists of two sub-systems (d'Alessandro, C. L., J., 1996):

(1). Text to phoneme's converter (Text to phoneme)

(2). Phoneme-to-speech converter (phoneme to speech)

The converter text to phoneme serves to change the input sentence in a particular language in the form of text into codes of sound, which is usually represented by the phoneme code, its duration and pitch.

Phoneme to Speech Converter functions to generate signals of speech based upon phoneme codes resulted from a previous process. This part largely involves diphone concatenation technique, which should be supported by diphone database comprising recorded speech segments such as a diphone (combination of two phoneme unit). Speech of a language is formed by a set of sound different from other languages. This is the primary reason why every language should have their own diphone databases.

There are several technical options in implementing such conversion. The most widely used are formant synthesizer and diphone concatenation. At present, the latter is more frequently used in application because it produces speech with natural quality. Diphone concatenation (Lenzo, Black et al., 2000) is a way of generating speech by combining sound segments consisting of two phonemes (diphone). To gain higher quality, some TTS used sound segment combination called multi-phone. The process of Text-to-Speech Conversion is shown in figure 1.

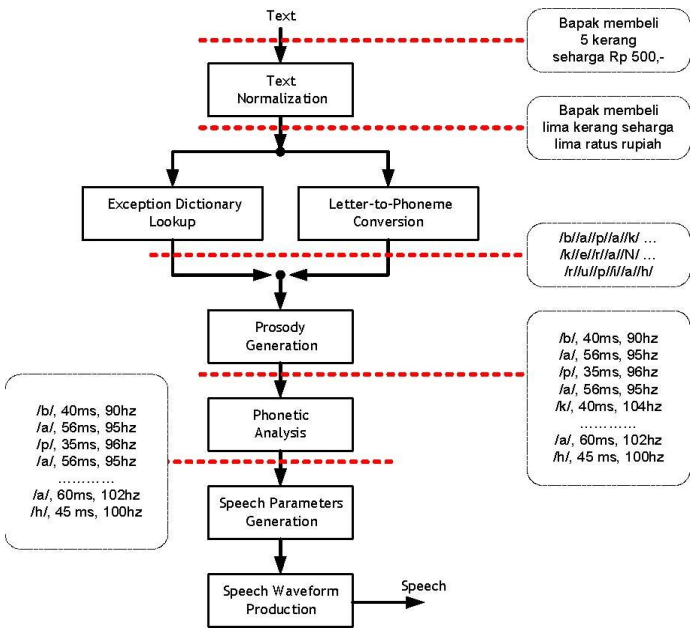

Figure 1. Diagram of text-to-speech the conversion process

According to (Lemmetty and Sami, 1999) TTS mechanism consists of two phases. The first is to analyze text in which text inputs are written or recorded as phonetic, linguistics or dialect. The second is to generate sound wave in which accoustic outputs are produced by the phonetic and prosody. The two phased mechanism is also called a high and low level combination synthesizer. A simpler version of the above procedure is represented in figure 2. Text inputs may be taken from the data sample of word processing, ASCII standard of email, SMS or text scanned from a newspaper. The string character is then reprocessed and analyzed. It results in the phonetic form where strings of phonemes with their intonation information, duration and appropriate stressing are concerned. Sound of words in the language is finally generated by a low-level synthesizer utilizing information from a previous synthesizing process (high-level synthesizer).

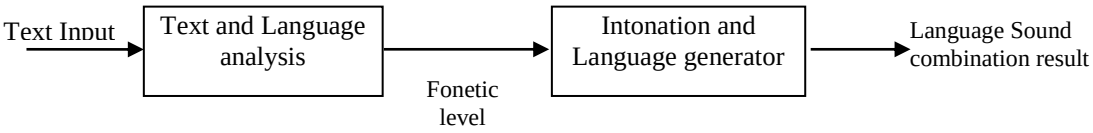

Figure 2. Simple application of TTS

DIPHONE CONCATENATIVE SYNTHESIS

Diphone Concatenative synthesis is one of the most popular methods used for creating a synthetic voice from recordings or samples of a particular person; it can capture a good deal of the acoustic quality of an individual, within some limits (Lenzo and Black, 2000). The rationale for using a diphone, which is two adjacent half-phones, is that the "center" of a phonetic realization is the most stable region, whereas the transition from one "segment" to another contains the most interesting phenomena, and thus the hardest to model. The diphone, then, cuts the units at the points of relative stability, rather than at the volatile phone-phone transition, where so-called coarticulatory effects appear.

There is clearly a simplifying assumption: that all relevant phonetic realizations can be enumerated, and by simply collecting all of phone-phone transitions, any possible sequence of speech sounds in the target language could be produced. Thus, with a 42-phone inventory, one could collect a 42 * 42 = 1764 diphone inventory and create a synthesizer that could speak anything, given the imposition of appropriate prosody – intonation, duration, and shift in spectral quality, as determined by othermodules in a general-purpose synthesizer.

DIPHONE

In a documented literature of MBROLA Project of the Faculté Polytechnique de Mons, TCTS, Belgium, diphone is stated as a language's sound unit starting from the middle where a phoneme is stable and ending at another middle of a phoneme that follows it. Diphone, or two phonems in a series are a component comprising of segments, which is a widely used component in TTS application (Dutoit and Leich, 1993). Some also say that diphone is a correlation between the phonemes (d'Alessandro, 1996).

In general, the number of diphones in a language is the multiple numbers of phonemes (Lemmetty and Sami, 1999). However, in daily conversation, there is usually some phonetic restriction (in pairs) which is not well formed. Such a difference exists in world languages. If tried, one can easily create what is called a non-existent diphone, and one of them has to always think about phoneme pairs through word limits. However, there is a certain case where a special

combination is hardly formed such as /h-h/. The diphone is not visible at the end of a syllable even though sometimes speakable when some aspiring sound (uttering a sound with an 'h' sound) is used to try to initiate a vocal.

DIPHONE DATABASE

In order to generate a language's sound a database of sounds with certain rules is needed. (d'Alessandro, 1996). Database is a computerized data-storage system that serves users of one or more organizations.

The idea behind building a diphone database, is obviously to make a list of possible phonetic transitions of a language is necessary (Lenzo, Black et al., 2000). That results in inaccuracy as well as becomes more practical, and It also simplifies assumption that articulating effects may reach no more than two phonemes (a diphone).

In regard to technical considerations and requested quality to achieve, the most optimum segment form is diphone or a set of two phonemes in aseries (Arman A.A, 2003). Method used for organizing diphones into speech is called diphone concatenation. The diphones are collected in a special database made with a special tool such as Diphone Studio 1.3. As shown in figure 3 and the database then becomes a complete collection of diphones.

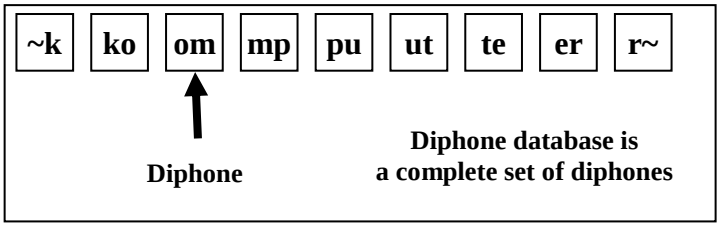

Figure 3. Common explanation of diphone concatenative

RESEARCH METHODOLOGY

DESIGN SYSTEM

Before the process of conversion of text into phonemes required beginning step is the normalization of text or text pre-processing as show figure 4. Text normalization is a process were written text is represented with spoken text compatible with human utterance. Before normalization: "This research began in 2005" After normalization: "This research began in two thousand and five." The last text is then converted to phonemes. Not all letters are similar with their phoneme codes. The function of conversion phoneme to speech is to generate signals of speech based upon phoneme codes resulted from a previous process.

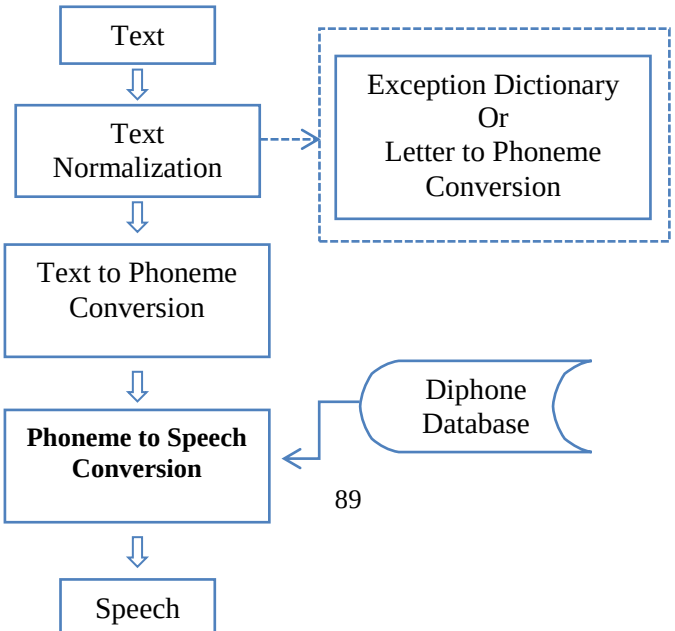

Figure 4. Diagram System Indonesian TTS

IMPLEMENTATION SYSTEM

The generally of steps, to create the system is:

(1) Developed Diphone database.
- Create a list of sample of words consisting of diphones organized by prioritizing looking diphone located in the middle of a word if not at the beginning or end.
- Recording the samples of words by segmentation
- Create diphones made with a tool Diphone Studio 1.3.

(2) Develop system using Microsoft Visual Delphi 6.0, includes:
- The conversion system from the input of numbers, acronyms, words, and sentences into representations diphone.

RESULTS AND DISCUSSION

There are 42 phonemes in Bahasa Indonesia with 10 pure vocals (monoftong), 3 double vocals (diftong), 28 consonants and a silent phoneme which added Up to 42 phonemes. Number of diphones produced is 42 X 42 = 1764 diphones. Speech synthesizing process on a sentence or a word is normalized and altered in a formula in which two phonemes are joined, or known best as diphone. All the diphones in the list are generated by a tool called Diphone Studio 1.3., a software that supports MBROLA. The result of joining two phonemes will be a diphone as shown by table 1.

Table 1. Sample of two joined phonemes into a diphone

No.	Fonem I	Fonem II	Difon	Sample Words
1.	[i]	[b]	/ i-b /	A*ib*
2.	[i]	[t]	/ i-t /	Pel*it*
3.	[i]	[k]	/ i-k /	Ba*ik*
4.	[i]	[u]	/ i-u /	H*iu*
5.	[e]	[o]	/ e-o /	B*eo*
6.	[e]	[j]	/ e-j /	*ej*aan, *ej*akulasi
7.	[e]	[k]	/ e-k /	*Ek*or
8.	[e]	[s]	/ e-s /	*Es*a
9.	[e]	[t]	/ e-t /	*et*os, *e*tika
10.	[ε]	[k]	/ ε-k /	nen*ek*, ej*ek*

In the Indonesian TTS system, there are three processes to follow, which are entering abbreviation data, converting text to phoneme and converting phoneme to speech. Entering an abbreviation, Data may be done by the user as one of the editing tools of abbreviation data where the description is saved in a table of abbreviation.

The process of converting text to phoneme is also useful for converting text input entered by the user. The text shall browse whether or not a marked element of abbreviation (written with capital letters and space) is available. If any, it shall continue to the abbreviation table. After converting text to phoneme is done, resulting in phonemes and are then converted to speech by combining two phonemes (a diphone) of the input text and searching the diphone in the diphone table. This phoneme to speech process results voices acceptable by users.

The conversion process from text to phoneme is processed by normalization of which functions to convert all sentence texts that will be uttered, into texts, which show how sentence should be uttered. For example, "Bapak membeli sepeda seharga Rp 250,000" (Daddy buys a bicycle for Rp250, 000). It is converted into speech text "Daddy buys a bicycle for two hundred and fifty thousand Rupiahs." Rp 250,000 is considered abnormal and should be normalized into speech sencente.

Of the word 'sepeda' there are phonemes of s,e,p,e,d and a. There are some phonemes with a similar letter but different pronounciation like "s-e" and "p-e" particularly ê phoneme is different from a phoneme, so that "sêpẽda" needs to be process through normalization with specific rules in this research to differenciate their pronounciation. To do this, a method of priority is used, searching for the highest priority. Users in turn may find specific words in the dictionary, put and convert them within this Indonesia TTS system.

Conversion may also occur in irregular or conditional environment depending upon neighboring letters or phonemes such as 'beda' which is different from 'bedak.' This also includes phonemes with similiar input letters but different pronounciation. There are also translational forms which their modus regubrity could not be found. To deal with such a case, this research uses the method of memory with the priority system. The priority rules out:

- There are only three priorities called priority 1, 2 and 3. In some diphones there are only two priorities, which are priority 1 and 2 this is based an how many phonemes are present, for instance, phonemes like e, ê, ẽ We are classified in priority 3 while U, u ; I,i ; n ñ ; o ô are classified in priority 1 and 2.
- The classification is carried out on similar phonemes of a word as like in the dictionary. Phoneme of e, ê, ẽ with diphones be, bê and bẽ in words like 'bebek, benar, benteng' where e in each word is pronounced differently, therefore the classification for the majority of words in a dictionary using phoneme ê in bê (benar) diphone are categorized in priority 1, and phoneme ẽ in diphone bẽ (benteng) with a similar amount of words are categorized in priority 2 and phoneme e in diphone be (bebek) with few similar amounts of words are grouped in priority.

The order of priority starts from 3 (least amount of words), if there is none, then continue to priority 2 and if there is none continue to priority 1. However, if a phoneme has only two priorities (1 and 2), then only priority 2 may take place if there is none priority 1 may take place. After normalization phase, a phonetic analysis may be conducted. This results in a series of diphones. After converting text to phoneme resulting in order of phonemes in a sentence or spoken word which are combined with two phonemes (diphones) such as in "Bapak" where there are three diphones; " _b; ba; pa; ak; ak; k_ ". The joining method is initiated with finding words or sentences where there is a phoneme with two letters like sy, kh, nk, ai, au, oi, ks, ny and ng. If found, the letters are, then grouped in a phoneme. For instance, in the word "sedang" there are

diphones "_s, se; ed; da; ang; ng_ " After joining phonemes are completed, then the method is continued by searching for the serial diphone in the diphone database already made. The final step (show figure 5) is generating the speech in form of sound acceptable and understood by users.

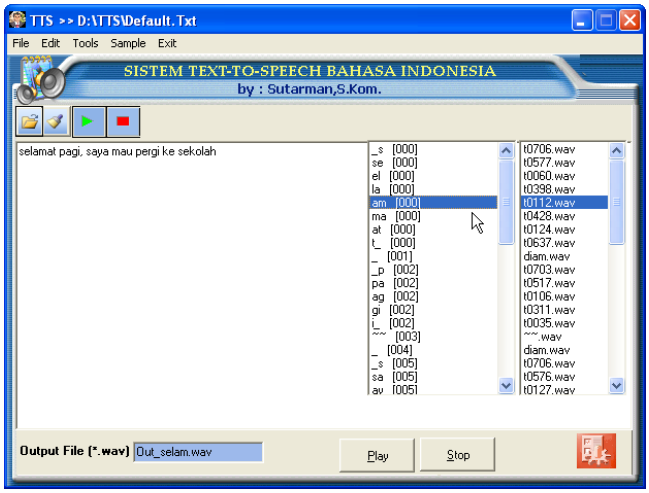

Figure 5. Implemention Indonesian TTS System

Converting text to phonemes is followed by converting them to pronounciation by joining two phonemes (a diphone) from the input text and searching them on the table of diphone. Converting phonemes to speech results in acceptable sound reqognized by users.

CONCLUSION

Indonesian Text-To-Speech System using diphone concatenative synthesis can produce speech or language the natural approach. Speech or resulted sound may not be perfect for several causes such as poor recording and distorted phoneme segmentation process where front, middle and end borders are not sincronized. The Indonesian Text to Speech system can differentiate special phonemes like in 'Beda' and 'Bedak' but sample of other spesific words is necessary to put into the system. This Indonesia TTS system can handle texts with abbreviation, there is the facility to add such words.

The main constraints encountered during the preparation of the diphone database and Text to Speech System, are: (1) the process of finding an example of the word in phoneme is less precise so the result is not perfect. (2) In the segmentation process diphone still much less precise, it is possible that the recording is not clear.

REFERENCES

Arman, A. A. 2002. Converting Text to phonemes. http://www.sgu.ac.id/library/ garuda/swf/IT/2010/Irfan.swf.

Arman, A. A. 2003. Building a database Diphone (MBROLA based). from http://lss-gtw.ee.itb.ac.id/~aa/indotts/diphone_dev.html.

Aulia A. 2012. Speech Optimization for Indonesian Text-To-Speech System, Graduate Theses, Institut Teknologi Bandung.

d'Alessandro, C. L., J. 1996. Synthetic Speech Generation. . Survey of the State of the Art in Human Language Technology: 4-10.

Dutoit, T. and H. Leich. 1993. mbr-psola : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database. Speech Communication 13.

George. Q. 2001. *The Learner's Dictionary of Today's Indonesian*. Sydney :Allen & Unwin, ISBN 1864485434.

Handi D. R. B. and Miftahul H. 2011. Text Pre-Processing of Text To Speech Synthesis System for Speak of Indonesian Language. Tesis on Instute of Technology Bandung, Indonesia.

Johannes Tan. 2009. Bahasa Indonesia: Between FAQs and Facts, http://www.indotransnet.com/article1.html.

Lemmetty and Sami. 1999. Review of Speech Synthesis, Helsinky University of Technology.

Lenzo, B., A. W. Black and K. A. Lenzo. 2000. Building voices in the Festival speech synthesis system.

Lenzo, K. A. and A. W. Black. 2000. Diphone Collection And Synthesis.

Marsono. 1999. Fonetik. Jogjakarta, Gadjah Mada University Press.

Nur Aziza Azis, R. M. Hikmah, T. V. Tjahja and A. S. Nugroho. 2011. Evaluation of Text-to-Speech Synthesizer for Indonesian Language Using Semantically Unpredictable Sentences Test: IndoTTS, eSpeak, and Google Translate TTS. International Conference on Advanced Computer Science & Information Systems, Indonesia, Universitas Indonesia.

Richard M. and Aulia A. 2013, Indonesian Text-To-Speech System Using Syllable Concatenation:Speech Optimization, 3rd International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME) Bandung, November 7-8.

Rommel, E. 2005. Aplikasi SMS dengan Text To Speech bahasa Indonesia pada Sistem Operasi Symbian untuk Tuna Netra, Tugas Akhir.

Soebardi. 1973. Learn bahasa Indonesia pattern approach. Yogyakarta, Kanisius.