Journal of Signal and Information Processing, 2014, 5, 80-88
Published Online August 2014 in SciRes. http://www.scirp.org/journal/jsip
http://dx.doi.org/10.4236/jsip.2014.53010

An Intonation Speech Synthesis Model for Indonesian Using Pitch Pattern and Phrase Identification

Yohanes Suyanto, Subanar, Agus Harjoko, Sri Hartati

Department of Computer Science and Electronics, Gadjah Mada University, Yogyakarta, Indonesia
Email: yanto@ugm.ac.id, subanar@ugm.ac.id, aharjoko@ugm.ac.id, shartati@ugm.ac.id

Received 25 May 2014; revised 20 June 2014; accepted 16 July 2014

Copyright © 2014 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Prosody in speech synthesis systems (text-to-speech) is a determinant of tone, duration, and loudness of speech sound. Intonation is a part of prosody which determines the speech tone. In Indonesian, intonation is determined by the structure of sentences, types of sentences, and also the position of the word in a sentence. In this study, a model of speech synthesis that focuses on its intonation is proposed. The speech intonation is determined by sentence structure, intonation patterns of the example sentences, and general rules of Indonesian pronunciation. The model receives texts and intonation patterns as inputs. Based on the general principle of Indonesian pronunciation, a prosody file was made. Based on input text, sentence structure is determined and then interval among parts of a sentence (phrase) can be determined. These intervals are used to correct the duration of the initial prosody file. Furthermore, the frequencies in prosody file were corrected using intonation patterns. The final result is prosody file that can be pronounced by speech engine application. Experiment results of studies using the original voice of radio news announcer and the speech synthesis show that the peaks of F_0 are determined by general rules or intonation patterns which are dominant. Similarity test with the PESQ method shows that the result of the synthesis is 1.18 at MOS-LQO scale.

Keywords

Speech Synthesis, PESQ, Intonation, Indonesian

1. Introduction

Modeling emotions in speech synthesis are made based on a number of parameters such as place, the level of the fundamental frequency (F_0), voice quality, articulation or accuracy. A different speech synthesis technique

How to cite this paper: Suyanto, Y., Subanar, Harjoko, A. and Hartati, S. (2014) An Intonation Speech Synthesis Model for Indonesian Using Pitch Pattern and Phrase Identification. *Journal of Signal and Information Processing*, **5**, 80-88.
http://dx.doi.org/10.4236/jsip.2014.53010

aimed for parameters in the different levels Schroder 2001. Formant synthesis, also known as the rule-based synthesis, creates the sound of speech based on the rules alone. No people's voice recording is involved in it. The result is a speech sound like a robot voice.

In the concatenation synthesis, voice recording from speaker strung together to produce synthetic speech. In this process, diphone which cuts the signal in mid phoneme until next mid phoneme is often used. Diphone is recorded with a monotonous tone. At the synthesis, required contour F_0 is constructed with signal processing techniques that result in a slight distortion. But the final result is more natural than formant synthesis [1]. In most diphone synthesis systems, only F_0 and duration can be controlled, while intensity for each segment is not easy to control.

Emotions of speaker such as neutral, joy, boredom, anger, sadness, fear, and indignation are reflected in the duration and intonation of speech from a person [2]. By copying pitch and duration of the original utterances to a monotonous one, it can be proved that both factors are sufficient to express various emotions. Research results show that emotion can be expressed simply by manipulating pitch and duration. Intonation is a set of syntactic prosody inherent in the utterance sentence [2]. The intensity of sound is not manipulated.

Contour tones containing duration and pitch are calculated by converting an intonation plan into a sequence of prosodic tone and highly dependent on the model of the speaker [3]. Contour patterns derived from the speaker's voice are applied at the time of speech synthesis.

Research on Indonesian speech synthesis which makes the coupling phoneme based speech synthesis has been done [4]. In general, the results can pronounce words in Indonesian quite fluently and can be understood by most listeners. However, speech synthesis has not produced intonation patterns (prosody) as the original speech [4]. Still there are some inaccuracies in the assembly of phonemes [4].

In Indonesian, utterance position on word stress does not depend on the number of syllables, but the stress falls on the penultimate syllable [5]. Prosodic type is characterized by F_0 and intensity. The pressure of nouns group on preverb position will be marked by the duration instead of the frequency. In this group, two peaks of F_0 are found.

2. Speech Synthesis Model Using Pitch Pattern and Phrase Identification

Speech synthesis is built to converts the input text into speech. Conversion of text into speech considering a sentence structure, intonation patterns, as well as a normalization. The results are in the form of text that qualifies as input of voice generator.

Prosodic rules which determine the frequency, duration, and intonation was compiled from the results of studying materials about intonation in Indonesian [6]. This rule is encoded to ease the implementation later. Speech synthesis parameters are grouped into four categories, namely pitch, duration, quality and articulation. The pitch parameter determines the value of F_0 [7]. The duration parameter determines the duration of rhythm control and rate of speech. Usually the pitch and duration parameters are associated with the phenomenon of linguistic (words or phrases) [3] [8]. The sound quality of speech is a parameter that determines the overall sound quality. The articulation parameter determines clearance of speech. Based on these parameters the model of speech synthesis is composed and shown in **Figure 1**.

The text comes from standard input or the file is read by the text normalization module in order to obtain text which is free of numbers and signs so that the whole text can be pronounced. The output of a normalization module text-norm will be used by the selector pattern, syntax analyst, and synthesis module. The selector pattern module will produce a pattern based on text-norm and an existing pattern database. This module is responsible for the pitch of each phoneme. Syntax analyst module will generate sentence syntax structure of the text-norm with the phrases and words interval duration information.

The synthesis module combined text-norm from normalization module, phonemes with prosody information in it based on the pattern from pattern selector, and attributes of sentence structure from syntax analysts into phonemes and prosody form. This module is responsible for the articulation and the quality of speech. Finally, this form will be voiced by DSP module.

2.1. Normalization Module

The text is a phrase that came from the standard input or files that may be contains numbers or punctuation

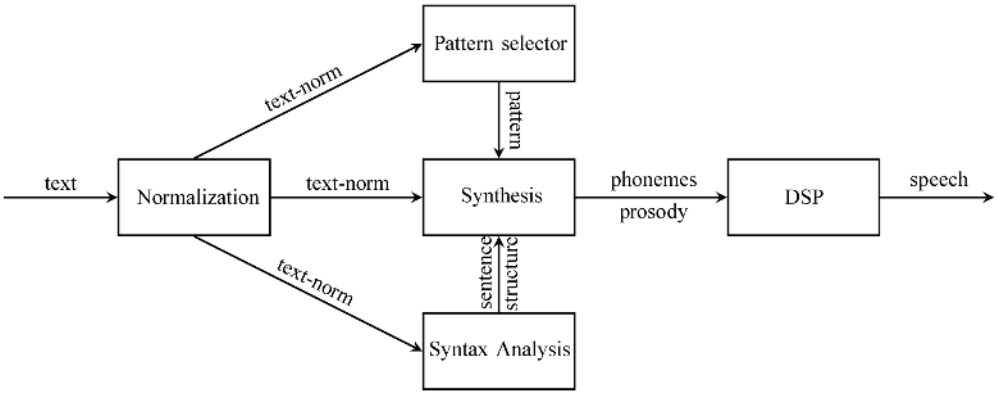

Figure 1. The proposed model of speech synthesis.

marks that can be pronounced. If there are numbers then this module will convert it into text. For example, in the text there is a number "12345" then it will be converted to "dua belas ribu tiga ratus empat puluh lima". Similarly, if there is an abbreviation, like "UGM" then it will be changed to "u-ge-em". Symbols need to be changed too, e.g. "=", ">", or "<".

2.2. Fundamental Frequency Pattern Module

Intonation patterns in the form of pairs of time t and fundamental frequency F_0 obtained by extraction F_0 of radio announcer voice. Based on the previously defined parameters the intonation pattern chose from a set of intonation patterns. The selection is done by paying attention on the pattern duration and length of the text. The duration of the speech text can be calculated to about sum of the duration of all phonemes. Results of pattern selector are F_0 for time from 0 to t. The pattern have been adapted to the text by interpolation method.

2.3. Syntactic Module

This module perform the *text-parsing* norm based standard grammatical structures. An important outcome of this module is the determination of the duration of phonemes and duration of pauses between phrases or between words.

Syntax analysis that will be used is Bahasa Indonesia structure rules [9]-[11]. The rules are converted into a context-free grammar structure (CFG). One notation is often used to write the CFG is Backus-Naur Form—BNF. Here are the example segments of Bahasa Indonesia BNF:

$$\langle \text{frasa} - \text{nominal} \rangle ::= \langle \text{frasa} - \text{nominal} \rangle \langle \text{frasa} - \text{nominal} \rangle. \quad (1)$$

$$\langle \text{frasa} - \text{nominal} \rangle ::= \langle \text{frasa} - \text{nominal} \rangle \langle \text{frasa} - \text{adjektival} \rangle. \quad (2)$$

$$\langle \text{frasa} - \text{nominal} \rangle ::= \langle \text{frasa} - \text{nominal} \rangle \langle \text{frasa} - \text{verbal} \rangle. \quad (3)$$

$$\langle \text{frasa} - \text{nominal} \rangle ::= \langle \text{frasa} - \text{nominal} \rangle \langle \text{penentu} \rangle. \quad (4)$$

$$\langle \text{frasa} - \text{nominal} \rangle ::= \langle \text{nominal} \rangle. \quad (5)$$

$$\langle \text{frasa} - \text{verbal} \rangle ::= \langle \text{verba} \rangle \langle \text{frasa} - \text{nominal} \rangle. \quad (6)$$

$$\langle \text{frasa} - \text{verbal} \rangle ::= \langle \text{verba} \rangle \langle \text{frasa} - \text{preposisional} \rangle. \quad (7)$$

$$\langle \text{frasa} - \text{verbal} \rangle ::= \langle \text{verba} \rangle \langle \text{frasa} - \text{nominal} \rangle. \langle \text{frasa} - \text{preposisional} \rangle. \quad (8)$$

Noun phrase ("frasa-nominal") can be construct from noun ("nomina") or noun phrase followed by another phrase. For example, noun phrase and verb phrase ("frasa-verbal"). In another segment verb phrase can be con-

struct from verb ("verba") and another phrase. Completed BNF for Bahasa Indonesia has been made.

2.4. Synthesis Module

Sentence in the text is converted into the format of pho with attention to intonation patterns and phoneme duration and duration of pauses between phrases. Intonation patterns derived from the module selector pattern, while the duration of phonemes and pause duration is derived from syntax analyst module. The pho format is ready to be fed to the DSP module to be voiced.

From pattern selector module resulting array of phonemes, includes spaces, and pho notation for the corresponding phonem. On the other hand, analyst module generates duration of pauses between words from the input sentence. The length of this pause is then used to correct the lag length of the pattern selector module output. It is expected the end result is in accordance with the intonation patterns and also in accordance with the structure of the sentence.

Illustration of the results of output from the intonation pattern module, the sentence structure analyst module, and synthesis module are shown in **Figure 2**. Output of intonation patterns module focused on intonation while sentence analyst module more focused on the pause duration between words.

3. Results

Fist, a pattern that contents of series of F_0 is got from a recorded voice by Praat application. Then, the input text was normalized by a normalized module written in PHP. Its output is called as text-norm. The analyst syntax running on text-norm gets the sentence structure with duration of phrases and words. The synthesis module composed a series of phonemes and prosody from the norm-text, the F_0 pattern, and the sentence structure. Finally, a voice will be generated by MBROLA application from this series of phonemes and prosody.

In this research some tools are used e.g. a Praat and an MBROLA applications. The Praat application is used to get F_0 from voice files while MBROLA application used to generate voice from phonemes + prosody form. MBROLA is not a speech synthesis complete software. MBROLA requires input of phonemes and prosody in a form that matches to the MBROLA form called pho. That is why the text to be synthesized must be converted first into MBROLA form (pho). MBROLA made by TCTS Lab of *Facult*? *é Polytechnique de Mons Belgium* [12].

Table 1 shows the result of normalization module. Symbols, numbers, and abbreviation converted correctly. The punctuation marks: comma $(,)$, period $(.)$, and hyphenation $(-)$ are still unchanged.

F_0 of sentence "*Empat LSM lembaga swadaya masyarakat diantaranya ICW indonesia corruption watch PSHK pusat studi hukum dan kebijakan menyampaikan aspirasi kepada panitia ad hock PAH empat DPD di*

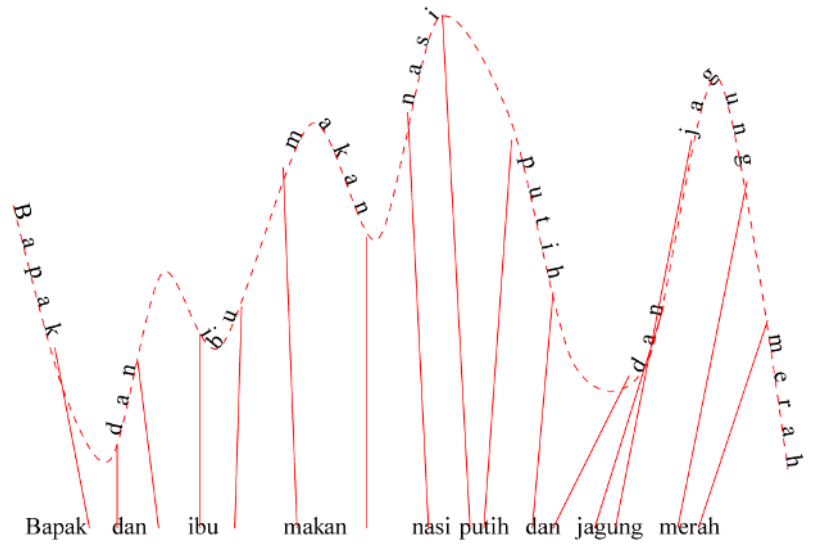

Figure 2. Relation between selector pattern and syntax analyst (sentence structure) modules. Top: result of intonation pattern, bottom: result of syntax analyst.

komplek parlemen senayan Jakarta hari ini" is shown in **Table 2** but there are only partial results. A full result contained about 1200 data. The F_0 table obtained by Praat application [7] [13].

If a sentence "*DPD kemudian tetap melakukan uji kepatutan dan kelayakan*" fed into the syntax analyst module, a sentence structure diagram obtained as shown in **Figure 3**.

Voice synthesis module in this study is part of the pho file processing module according to the rules of normalization followed the general tone combined with syntax analysts and intonation patterns. For comparison, the results will be presented waveform synthesis with flat intonation, intonation with the general rule, and intonation pattern.

Table 1. Result of normalization module for complete sentences.

Before	After
Pemerintah menyatakan ekspor sepanjang tahun **2013** secara kumulatif melampaui target.	pemerintah menyatakan ekspor sepanjang tahun **dua ribu tiga belas** secara kumulatif melampaui target.
"Target ekspor **2013** itu **179** miliar dollar AS, dengan angka Desember, kumulatif kita mencapai **182,57** miliar dollar **AS**," kata Wakil Menteri Perdagangan Bayu Krisnamurthi, Senin	"target ekspor **dua ribu tiga belas** itu **seratus tujuh puluh sembilan** miliar dollar **a es**, dengan angka desember, kumulatif kita mencapai **seratus delapan puluh dua koma lima tujuh** miliar dollar **a es**," kata wakil menteri perdagangan bayu krisnamurthi, senin
Hal itu berdasarkan pengumuman Badan Pusat Statistik (**BPS**).	hal itu berdasarkan pengumuman badan pusat statistik **be pe es**
KJRI terus memantau kondisi kesehatan **ybs**., dengan melakukan pemeriksaan medis secara rutin ke rumah sakit.	**ka je er i** terus memantau kondisi kesehatan **yang bersangkutan**, dengan melakukan pemeriksaan medis secara rutin ke rumah sakit.
Berdasarkan hasil pemeriksaan, saat ini **Sdri.** Kokom dalam kondisi sehat, dan luka-lukanya telah berangsur-angsur pulih serta telah dapat berjalan secara normal, meskipun masih mengeluhkan nyeri dibagian kaki kirinya.	berdasarkan hasil pemeriksaan, saat ini **Saudari** kokom dalam kondisi sehat, dan luka-lukanya telah berangsur-angsur pulih serta telah dapat berjalan secara normal, meskipun masih mengeluhkan nyeri dibagian kaki kirinya.

Table 2. Excerpts F_0 results for sentence.

Time (s)	F_0 (Hz)
0.622125	138.447701
0.632125	134.561068
0.642125	--undefined--
0.652125	--undefined--
0.662125	--undefined--
0.672125	--undefined--
0.682125	--undefined--
0.692125	--undefined--
0.702125	--undefined--
0.712125	--undefined--
0.722125	151.699538
0.732125	149.879256
0.742125	149.277178
0.752125	149.884325
0.762125	151.686222
0.772125	153.601689
0.782125	154.488904
0.792125	152.781495

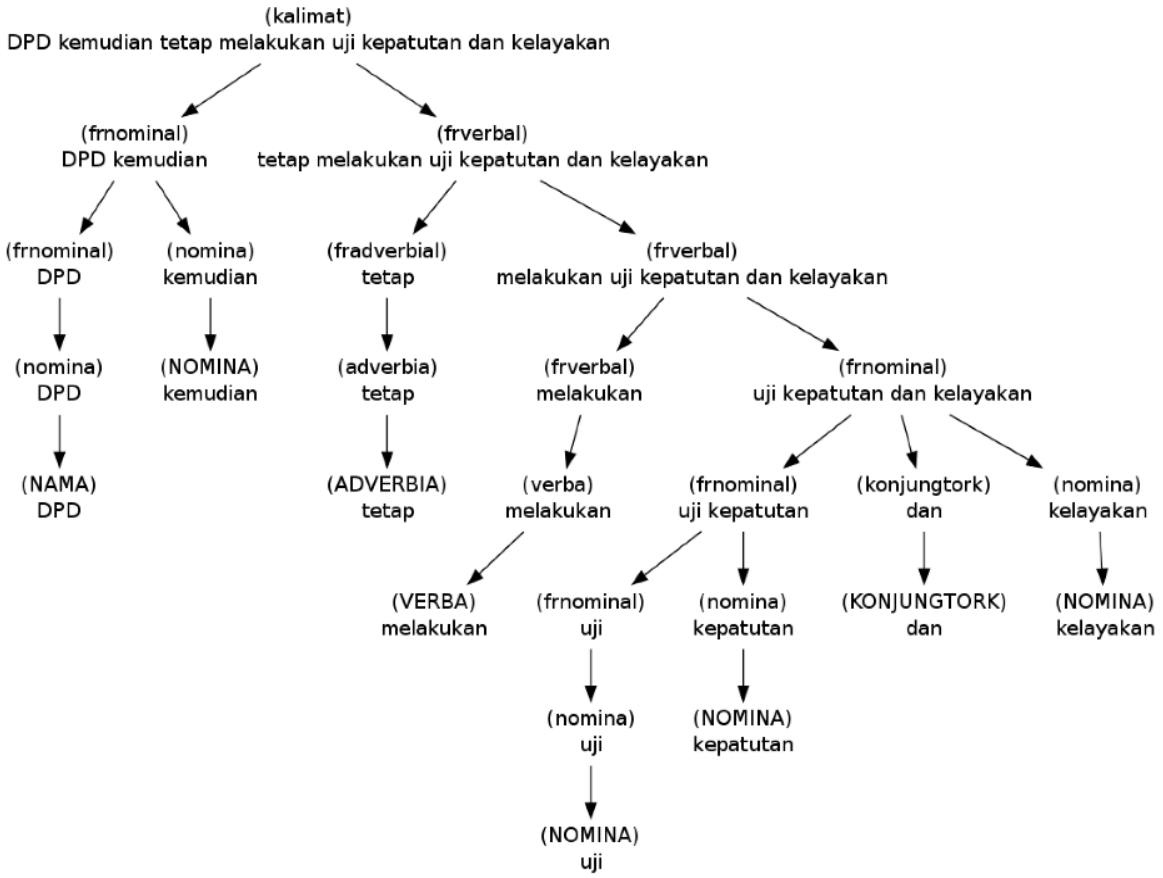

Figure 3. Example of sentence structure.

Firstly, the sentence "*Wimar tidak dapat mengikuti pemakaman Gus Dur di Jombang*" is synthesized with flat intonation. The results of the graph F_0 are shown in **Figure 4**. Secondly, for the same sentence but intonation synthesized according to the general rules, the results of the graph F_0 are shown in **Figure 5**. And, **Figure 6** shows the synthesis of the same as the second but improved by incorporating elements of intonation pattern. The selected intonation pattern has F_0 as shown in **Figure 7**.

Similarity checking beetween synthesis result and original voice can be done by quantity approach. In this study, the checking similarity uses the PESQ method [14] [15]. The result of PESQ for several sentences is shown in **Table 3**.

4. Discussion

Table 1 shows that the integer number of year (2013) and the amount of money (179) can be converted correctly. Likewise fractions (182,57) can be converted to text correctly. Because 57 is a fraction, so it is converted into a "lima tujuh" (five seven) instead of "lima puluh tujuh" (fifty seven). The comma in 182,57 is decimal separator instead of thousand separator. Abbreviations and acronyms can be converted correctly. KJRI, BPS, and AS are converted to "ka je er i", "be pe es", and "a es". The "ybs." and "Sdri." acronyms are converted to "yang bersangkutan" and "Saudari".

Praat application output has 2 columns *i.e* time and fundamental frequency (F_0). In this study, the time has 0.01 seconds interval so there are 100 data in every second of voice. See **Table 2**, a graph like **Figure 7** can be obtained from this table. The graph shows that it has 2 peaks in 9^{th} second and 15^{th} second. During the first 7.5 seconds, there are several small peaks. Those peaks can be dominated when combination of general rules and intonation patterns has made.

A flat intonation was obtained from MBROLA when the frequency was set in one value for a long of speech

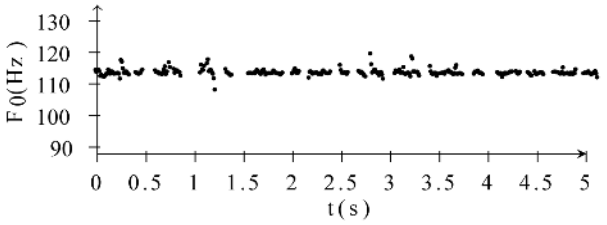

Figure 4. Graph F_0 with flat intonation.

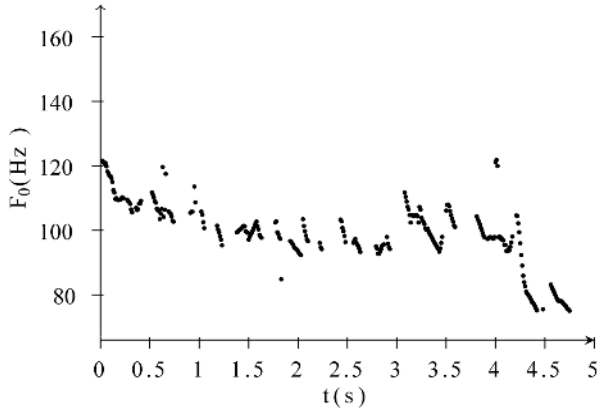

Figure 5. Graph F_0 with intonation according to general rules.

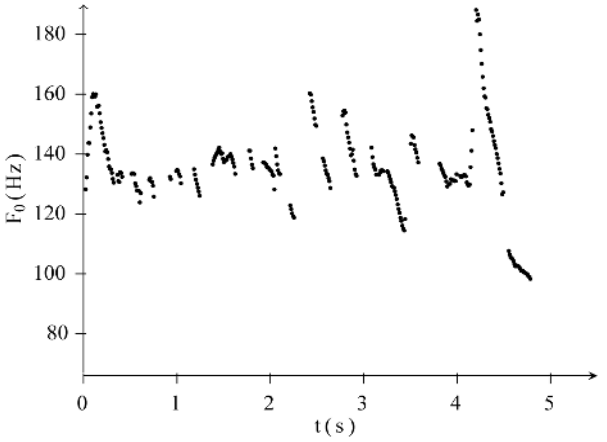

Figure 6. Graph F_0 with the general rules and intonation patterns.

synthesis. **Figure 4** shows the result of flat intonation graph when the frequency set to 115 Hz. Although several small peaks were shown in a flat intonation, it happened because of MBROLA characteristic.

In gerenal rules of intonation according to [5] the peaks of F_0 appeared at the syllables of nouns. For nouns before a verb F_0 peak occur in a syllable before the last syllable called the penultimate. In a sentence "*Wimar tidak bisa mengikuti pemakaman Gus Dur di Jombang*" "Wimar" is a noun before the verb "mengikuti". In that case F_0 peak occur in syllable "Wi" from the word "Wimar". That is in the start of sentence. See the first seconds of graph in **Figure 4**. For nouns after a verb F_0 peak occur in the first syllable of the first noun. In the sentence the first syllable of the first noun is "Gus". This syllable is the 15^{th} syllable of 19 syllables in the sentence. We can calculate that this syllable occur in about

$$t = \frac{15}{19} \times 5 \text{ s} = 3.9 \text{ s} \tag{9}$$

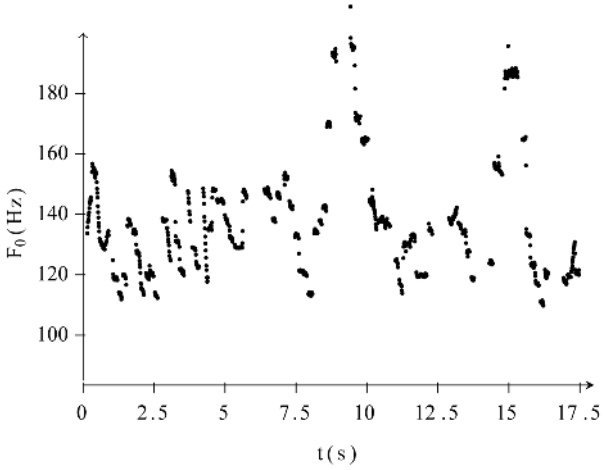

Figure 7. Graph F_0 of selected pattern.

Table 3. Values of PESQ for several sentences.

Reference	Synthesis	PESQ
kal101.wav	kal101-sin.wav	1.272
kal102a.wav	Kal102a-sin.wav	1.124
kal103a.wav	kal103a-sin.wav	1.067
kal104.wav	kal104a-sin.wav	1.048
kal105.wav	kal105-sin.wav	1.062
kal106.wav	kal106-sin.wav	1.037
kal107.wav	kal107-sin.wav	1.064

This peak can be seen at the fourth second of the graph in **Figure 5**. After that the values of F_0 continues to fall until the end of the sentence.

The pattern has F_0 graph as shown in **Figure 7**. It has two peaks that occur at 9^{th} and 15^{th} second from 17.5 s speech synthesis a long. We can calculate the segments of its value

$$\frac{9}{17.5} = 0.51 \tag{10}$$

$$\frac{15}{17.5} = 0.86 \tag{11}$$

This selected pattern then combined to speech voice that has F_0 as **Figure 5** obtained a speech voice that has F_0 as **Figure 6**. The F_0 peaks of the syntesis result can be predict from Equation (9) occur at

$$0.79 \times 5 \text{ s} = 3.9 \text{ s}$$

This results is not match with the graph second peak in **Figure 6**. There is no peak in this value. If we used Equation (10) and (11) we get

$$0.51 \times 5 \text{ s} = 2.55 \text{ s}$$

and

$$0.86 \times 5 \text{ s} = 4.3 \text{ s}$$

There are peaks that match with peaks in **Figure 6**. The second peak can take a place the wrong result as Eq-

uation (9). Thus **Figure 6** is combination of **Figure 5** and **Figure 7**.

From **Table 3**, we get that the average value of PESQ is 1.096. This value can be converted to MOS-LQO scale by the following formula [14]

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945 * x + 4.6607}}$$

where y is MOS-LQO in the range 1.02 to 4.56, x is PESQ value in the range -0.5 to 4.5. The value of MOS-LQO is

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945 * 1.096 + 4.6607}} = 1.18$$

5. Conclusion

Speech synthesis model with a combination of general rules and intonation patterns of the fundamental frequency (F_0) can be implemented. The peaks of F_0 determined by general rules or intonation patterns are dominant. Indonesian has intonation general rules, but should not be adhered strictly. Similarity test with the PESQ method shows that the synthesis results are about 1.18 at MOS-LQO scale. A speech synthesis based on the announcer voice intonation patterns is expected to be used as an alternative method of Indonesian speech synthesis refinement.

References

[1] Schröder, M. (2001) Emotional Speech Synthesis: A Review. *Proceedings of Eurospeech* 2001, **1**, 561-564.

[2] Vroomen, J., Collier, R. and Mozziconacci, S. (1993) Duration and Intonation in Emotional Speech. *Proceedings of the Third European Conference on Speech Communication and Technology*, Berlin, 22-25 September 1993, 577-580.

[3] Campbell, W.N., Isard, S., Monaghan, A.L.C. and Verhoeven, J. (1990) Duration, Pitch and Diphones in the CSTR TTS System. *Proceedings of the International Conference on Spoken Language Processing*, Kobe, 1 January 1990, 825-828.

[4] Tritoasmoro, I.I. (2006) Text-to-Speech Bahasa Indonesia Menggunakan Concatenation Synthesizer Berbasis Fonem. *Seminar Nasional Sistem dan Informatika*, Bali, 17 November 2006, 171-176.

[5] Laksman, M. (1995) Realisasi Tekanan Kata dalam Bahasa Indonesia. PELLBA 8, pages 179{215. Lembaga Bahasa Unika Atmajaya, Jakarta, 1995.

[6] Halim, A. (1975) Intonation in Relation to Syntax in Bahasa Indonesia. Proyek Pengembangan Bahasa dan Sastra Indonesia dan Daerah, Departemen Pendidikan dan Kebudayaan.

[7] van Lieshout, P.P. (2003) PRAAT Short Tutorial. University of Toronto, Graduate Department of Speech-Language Pathology, Faculty of Medicine, Oral Dynamics Lab (ODL), Toronto.

[8] Heuven, V.J.V. and Zanten, E.V. (2007) Prosody in Indonesian Languages. LOT, Utrecht.

[9] Sakri, A. (1994) Bangun Kalimat Bahasa Indonesia. 2nd Edition, Penerbit ITB, Bandung.

[10] McCune, K.M. (1985) The Internal Structure of Indonesian Roots. Number v. 2 in the Internal Structure of Indonesian Roots. Badan Penyelenggara Seri Nusa, Universitas Katolik Indonesia Atma Jaya, Jakarta.

[11] Vamarasi, M.K. (1986) Grammatical Relations in Bahasa Indonesia. Cornell University, Ithaca.

[12] Mbrola, T. (2009) The MBROLA Home Page. http://tcts.fpms.ac.be/synthesis/mbrola/

[13] Bořil, H. and Pollák, P. (2004) Direct Time Domain Fundamental Frequency Estimation of Speech in Noisy Conditions. *EUSIPCO* 2004, **2**, 1003-1006.

[14] ITU (2001) ITU-T Recommendation P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. Technical Report, ITU.

[15] Rix, A.W., Beerends, J.G., Hollier, M.P. and Hekstra, A.P. (2001) Perceptual Evaluation of Speech Quality (PESQ)—A New Method for Speech Quality Assessment of Telephone Networks and Codecs. *IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, Salt Lake City, 7-11 May 2001, 749-752.

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.