PROSODY MODELING OF IMPERATIVE SENTENCE IN BAHASA INDONESIA USING MATLAB

Arum Tri Iswari Purwanti
Universitas Gunadarma
Email: arum_tri@staff.gunadarma.ac.id

Muhammad Subali
Sekolah Tinggi Teknik Multimedia Cendekia Abditama (STTC)
Email: muhammadsubali@yahoo.com

ABSTRACT

Speech synthesis or text to speech is a speech processing technology and can change text to speech form. Speech synthesis applications are also available in Bahasa Indonesia which is known as TTS Indo. In these applications, there are still many shortcomings such as the sound produced using a sentence with a flat tone (without prosody), whereas in a sentence that prosody used to distinguish the meaning of the sentence. The sentences which are discussed in this journal are imperative sentences. Imperative sentence modeling on this research is basically seeing pitch (fundamental frequency) of the speaker by using MATLAB software. By looking at the results of the imperative sentence for Bahasa Indonesia, either one word or two words, both have a high tendency of intonation in the final seconds (at the end of the sentence). At the last, the prosody modeling analysis that inform the alteration of pitch can be used for prosody model of TTS Indo.

Key words: Pitch, Bahasa Indonesia, Imperative Sentence, Prosody Modeling.

Introduction

A. Introduction

Text-to-speech is widely used for all people in this world. Text-to-speech or which is known as speech synthesis is an application that converts an input of text and output of speech. Generally the use of speech synthesis is the text pronunciation aided for people who have difficulty in speaking. A simplified version of the procedure is presented in Figure 1. Improvement quality of speech synthesis system has been more remarkable. The system can produce output that sounds relatively close to human language. For the ears, two more important elements highlighted in this regard are the improvement of signal quality and improvement in the coherence and naturalness. The improvements in signal quality of good contemporary systems are mainly due to the use and improved control over concatenative speech technology, while the greater coherence and naturalness of synthetic speech are primarily a function of much improved prosodic modeling [Keller, 2002].

Figure 1. Simple text-to-speech synthesis procedure [Lemmetty, 1999]

Converting input of text into output of speech relies heavily on the rules which apply in a language. Language itself is a tool used by humans to be able to communicate with other humans. Language has a characteristic that also depends on one’s area of origin. Not only where the area of someone is, but also how a person uttering a word also helped determine. The symbols are inputted text will naturally be represented in sound in a language [Arman, 2008]. Now, high-quality speech synthesizer has been available for some languages, e.g. English, French, Dutch, German and several other languages. Speech synthesis applications are also available in Bahasa Indonesia, where an application is often known as Indo TTS developed by Arry Akhmad Arman. But after trying this application there are still many shortcomings such as the sound produced using a sentence with a flat tone (without prosody). Whereas in a sentence has prosody which is used to distinguish the meaning of the sentence [Novianti, 2009].

The objective of this research is to analyze the intonation (prosody) of Imperative Sentences for Bahasa Indonesia. By analyzing the intonation of Imperative Sentences one by one, it will get the prosody modeling of the Imperative Sentences on Bahasa Indonesia. In the process of analyzing the Imperative Sentences used MATLAB Simulation
Software. The use of this software is to help obtaining the pitch values for every sentence, then processing it into a prosody form of Imperative Sentences on Bahasa Indonesia.

In this paper, we describe Bahasa Indonesia especially Imperative Sentence and intonation modeling as supporting theoretical. For the methodology research, Imperative Sentence has been chosen to define the prosody. It continued by recording each sentences which has samples and processing of data analysis. The main components of this research are the pitch obtaining process, filter definition, and fundamental frequency estimation. In result and analysis, the data obtained in Bahasa Indonesia are one word and two words of Imperative Sentence. Each data sentences has each samples which is continued by processing the data synchronization and data sampling. Thus, the conclusion declare the imperative sentence in Bahasa Indonesia tends to have high intonation at the end of the sentence and this data can be used to further research to enrich the language library especially in the prosody of Bahasa Indonesia.

B. Review Literature

Languages use pitch variation contrastively for the expression of discoursal meaning and for marking phrases. One of the important points developed in is that intonation is structural [Ladd, 2008], just as lexical tone is structural, or morphological paradigms are. In principle, an intonation contour has two structures: a morphological one, which identifies the morphemes and thus gives the meaning of the contour; and a phonological one, which gives its tones [Gussenhoven, 2004] [Clark, 2003].

Most of the predicate that states the order of verbs or verbal phrases, it is usually not accompanied by the subject [Sugono, 1999]. In other words, all verbs can express command. One of the command verb features has not a beginning. Another feature of Imperative Sentence is that if it be written at the end of sentence with an exclamation mark (!) [Indonesia, 2000].

1. Tembak!
2. Catat semua keterangan saksi!
3. Dengar baik-baik!

In addition to such basic verbs, derived verbs can express command. In this case also the prefix me(N)- which states actively uninstalled [Irwan, 2009].

1. Perbaiki mobil itu!
2. Bersatu!
3. Bukukan kisah itu!

In addition, the predicate that states the command is marked also by the particle -lah. Commands that use these particles feel more assertive, as shown in the following example.

1. Bacalah buku!
2. Tuntutlah ilmu!
3. Berdoalah!

In reality it is often refined form of the command so that it becomes an invitation, request, or prohibition. This type is usually preceded by words such as mari, mohon, silakan, harap, or jangan.

The related research has been discussed in Prototype Prosody in Speech Disorder Children. [Subali and friends, 2013] This study used the Dynamic Time Warping (DTW) method to calculate the distance between two time series data of the sound signal. Testing of these applications is done by finding the ratio error matching that state the probability of matching errors in the system. In another research, Analisis Frekuensi Dasar dan Frekuensi Formant dari Fonem Huruf Hijaiyah untuk Pengucapan Makhraj dengan Metode DTW has been discussed. The sound will be extracted to get the value of the fundamental frequency and formant frequency. After getting both frequencies, it will be obtained analysis of the similarities and differences in the fundamental frequency and formant frequencies of speech beginner and expert and it will shows matching distance of both speech. The result is all of speech beginner and expert based on makhraj pronunciation have different values of fundamental frequency and formant frequency.

C. Methodology

Imperative Sentence modeling on this research is basically seeing pitch of the speaker. Pitch has been described in the previous chapter which is the auditory sensation of tonal height. A collection of pitch is then processed to obtain a prosody modeling of the words entered by the speaker. The Prosody Modeling Research Process Diagram of Imperative Sentence can be shown in figure 2.

The recording process begins by making the Imperative Sentence which is used as prosody modeling. As remembered the Imperative Sentence can be made only by a predicate word, the sentences for one word are as follows [Bahasa, 2008]:

1. Baca!
2. Catat!
Furthermore, the sentences for two words of Imperative Sentence used a refined form. A refined form has another purpose in the form of an invitation, request, or prohibition. This type of sentence is preceded by words such as "mari, mohon, silakan, harap, or jangan". Then the sentence used for this study are as follows:

1. Ayo baca!
2. Ayo makan!
3. Harap tenang!
4. Jangan lompat!
5. Jangan pergi!
6. Jangan tembak!
7. Mari duduk!
8. Minum!
9. Pergi!
In addition to hardware, this research also uses software to help the recording process (figure 3). Software that supports this research is the SFS/WASP version 1.5. WASP is a free program for the recording, display and analysis of speech. It can record and replay speech signals, save them and reload them from disk, edit annotations, and display spectrograms and a fundamental frequency track which is also designed to be compatible with the Speech Filing System (SFS) tools for speech research [Novianti, 2009].

Analyzing the data has been obtained aiming to get prosody modeling of the Imperative Sentence. This process is performed after executing the formatting process. The data analysis processes consisted of pitch obtaining process, data synchronization process, data sampling process, and produce the desired prosody modeling in Figure 4.

Pitch Obtaining Process
Pitch obtaining process, which is known as the estimation of fundamental frequency, or pitch, is an essential component of many speech processing applications [Gonzalez and Brookes, 2011]. The proposed method used in this research is adapted of “A Pitch Estimation Filter Robust To High Levels Of Noise (PEFAC)”. For a perfectly periodic source at frequency f0, our signal model at time t in the power spectral density domain is

\[
Y_i(f) = \sum_{k=1}^{K} a_{k,i}(f - kf_0) + N_i(f)
\]  

(1)

where \(N_i(f)\) represents the power spectral density of the unwanted noise and \(a_{k,i}\) the power of the kth harmonic. In the log-frequency domain, the signal model can be expressed as
where \( q = \log f \). In this domain, the spacing of the harmonics is independent of \( f_0 \) and their energy can therefore be combined by convolving \( Y_t(q) \) with a filter with impulse response

\[
Y_t(q) = \sum_{k=1}^{K} \alpha_k \delta(q - \log k - \log f_0) + N_t(q)
\]

The convolution \( Y_t(q) * h(q) \) will include a peak at \( q_0 = \log f_0 \) and additional peaks corresponding to simple rational multiples of \( f_0 \).

**Filter Definition**

In practice, the width of each harmonic peak will be broadened due to the analysis window and the rate of change of \( f_0 \). Accordingly we use a filter with broadened peaks having the impulse response

\[
h(q) = \beta \gamma \log(\gamma - \cos(2\pi \frac{q}{\gamma}))
\]

for \( \log(0.5) < q < \log(K+0.5) \) and \( h(q) = 0 \) otherwise. \( \gamma \) is an algorithm parameter that controls the peak width while \( \beta \) is chosen so that \( \int h(q) dq = 0 \). The number of peaks, \( K \), is restricted to 10 in order to reduce the response of \( Y_t(q) * h(q) \) at values of \( q \) corresponding to sub-harmonics of \( f_0 \).

**Fundamental Frequency Estimation**

The complete PEFAC (Pitch Estimation Filter with Amplitude Compression) therefore comprises the following steps whose outputs are shown in figure 5 for a single voiced frame corrupted by car noise:

1. Transform the input signal to the time-frequency domain using the short-time Fourier transform (STFT), \( Y_t(f) \)
2. Interpolate the power spectral density (PSD) of each frame onto a log-spaced frequency grid, \( Y_t(q) \)
3. Find \( a_t(q) \) so that the normalized smoothed spectrum \( Y'_t(q) \) equals \( L(q) \) and calculate the compressed PSD, \( Y'_t(q) \)
4. Convolve the compressed PSD, \( Y'_t(q) \), with the analysis filter, \( h(q) \), and select the highest peak in the feasible range as the estimated pitch.

Figure 5. Algorithm processing steps for a single voiced frame of speech corrupted with car noise. (a) PSD in dB, (b) PSD in dB in a log-frequency grid, (c) compressed PSD in dB in a log-frequency grid, and (d) normalized output of the filter and fundamental frequency, \( f_0 \).

**D. Result And Analysis**

After getting the raw data of 10 times (pronunciation) in a time interval of 2 seconds, it continued by splitting the data into 10 different data. Each data (samples) used as input in Matlab to seek the pitch (fundamental frequency). Some figures below show sample data of Imperative Sentence.

**The data obtained of Imperative Sentence in Bahasa Indonesia for two words**

The data obtained of Imperative Sentence in Bahasa Indonesia for one word obtain time of 0.7553 seconds. The data obtained have not been carried out the data synchronization so that data obtained are not uniform (due to data separation). It seen that the frequency of the sample 1 to sample 10 does not have the same frequency at first. For that, emphasized the importance of synchronization for the raw data. Data synchronization is a data analysis based on the distance (interval) and time.

Figure 6. Data Sample 1 of word "Ayo Baca"
Data Synchronization Process
As already explained, the data analysis based on distance and time is a data synchronization process to obtain absolute valued. After the data synchronization process is expected that the resulting data can be directly used as the primary data for further analysis. The record data that have been processed are analyzed to determine the length/distance and determine the sample points based on the time and frequency. Data synchronization is done to equalize the initial frequency to the time. Before the data synchronization process, data is very irregular. This occurs due to the data separation done before. Then this data synchronization is to smoothing the irregular data so that the next analysis can proceed. In other words, data synchronization is done so that the data obtained more "apple to apple" so it can be processed to obtain a most appear value of the frequency for one unit of time.

Data Sampling Process
After executing the sampling process, it followed by finding the average value of sampling data. There are 8 Imperative Sentence of one word and 7 Imperative Sentence of two words that have previously been through the stage of the sampling process. The results of the sampling process have a prosody modeling for each Imperative Sentence.

Prosody modeling is a data collection of frequency was selected after performing the sampling process. Prosody modeling of each Imperative Sentence is collected to find the average value. Collection of prosody modeling of every Imperative Sentence aims to seek a prosody modeling to the Imperative Sentence that has been used as input. In other words, for imperative sentence of one word and two words, the result of prosody modeling is different each other. Prosody modeling is the average data set of each Imperative Sentence for Bahasa Indonesia. The Figure 8 show a data set of each Imperative Sentence.

Table 1. Raw Data Sample of word "Ayo Baca"

<table>
<thead>
<tr>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Sample 1</td>
<td>Sample 2</td>
<td>Sample 3</td>
<td>Sample 4</td>
<td>Sample 5</td>
<td>Sample 6</td>
<td>Sample 7</td>
<td>Sample 8</td>
<td>Sample 9</td>
</tr>
<tr>
<td>6</td>
<td>0.0529</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
</tr>
<tr>
<td>7</td>
<td>0.0593</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
</tr>
<tr>
<td>8</td>
<td>0.0601</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
</tr>
<tr>
<td>9</td>
<td>0.0780</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
</tr>
<tr>
<td>10</td>
<td>0.0830</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
<td>125.4560</td>
</tr>
</tbody>
</table>

Table 2. Data collection of two words in Imperative Sentence
The processes carried out in this chapter are also done on other Imperative Sentence. As already known before, this research uses 8 one word and 7 two words. The sentences for one word are as follows, "Baca!, Catat!, Duduk!, Jalan!, Makan!, Minum!, Pergi!, Tembak!", and then the sentence used for two words are as follows, "Ayo baca!, Ayo makan!, Harap tenang!, Jangan lompat!, Jangan pergi!, Jangan tembak!, Mari duduk!".

Figure 7. (a) Prosody Modeling of one word in Imperative Sentence (b) General Prosody Modeling of two words in Imperative Sentence

E. Conclusion

The research that has been done, it started by learning literature study, collecting the Imperative Sentence for only one word and two words, processing the data recording, preparing the recording equipment, processing the record data formatting, and finally analyzing the record data. In analyzing the data, some of which stage of the process to obtain the pitch (fundamental frequency), the data synchronization process, the sampling data process included calculating the average value of sampling data, and finally to the ultimate goal of prosody modeling for one word and two words. By looking at the results of the Imperative Sentence for Bahasa Indonesia in the previous chapter, either one word or two words, both have a high tendency of intonation in the final seconds. This indicates the Imperative Sentence that tends to have high intonation at the end of sentence. Emphasis on the Imperative Sentence which written frequently by an exclamation mark (!) show that the presence of a high intonation on the sentence.

For further research, this data can be used for implementing Imperative Sentence in Speech synthesis applications in Bahasa Indonesia, where an application is often known as Indo TTS. Exclamation mark can be distinguish the Imperative sentence itself. When exclamation mark is used after the word, the pronunciation of the sentence should be changed as following this final pitch data.
References


