Chunking Phrase to Predict Pause Break in Pontianak Malay Language

Arif Bijaksana Putra Negara[a1], Yulia Magdalena [a2], Rudy Dwi Nyoto[a3], Herry Sujaini[a4]

[a]Informatics Department, Tanjungpura University
Prof.Dr.H.Hadari Nawawi Street, Pontianak, Indonesia
[1]arif.bpn@informatika.untan.ac.id
[2]ymyuliamagdalena@gmail.com
[3]rudydn@informatika.untan.ac.id
[4]herry_sujaini@yahoo.com

Abstract

Pause break is one of the indicators of speech to be easily understood in the Text-to-Speech System. This research aims to improve the accuracy of pause prediction in Pontianak Malay Language Sentences based on earlier research using a chunking phrase. This research is done as one of the efforts to preserve Pontianak Malay Language in order not to become extinct as a local language. Chunking method uses RegexpParser function in Natural Language Toolkit to crop sentences into phrases based on the Part of Speech type. In this research, the authors have developed a new grammar and pause break rule that is different from the earlier research to increase the accuracy of pause prediction. The data used is 500 Pontianak Malay Language sentences that have been recorded by a Pontianak Malay Language native speaker to get the pause break analysis. The pause consists of a short pause (symbolized as "/1) and a long pause (symbolized as "/2"). The tests were a test of pause break compatibility in one sentence and a test using f-measure, recall, and precision parameters. Based on the tests that have been done, the new grammar rule and pause break rule from this research have a better prediction accuracy than the earlier research with the correct predictive value of sentences increasing by 23% from the earlier rule.

Keywords: *Pause Break, Chunking, Grammar Rule, Pause Break Rule, Accuracy, Text-to-Speech, Pontianak Malay*

1. Introduction

A language is a communication tool used in human life. In Indonesia, besides Indonesian as the national language, there are many languages born and developed in certain regions and are called local languages. Pontianak Malay language is a Malay dialect spoken by the people of Pontianak City, Kubu Raya Regency, and Mempawah Regency and has similarities with Malay Peninsula Malay (Johor-Riau) [1]. This language has been used as communication tools in Pontianak. Based on the results of the population census conducted by Statistics Indonesia, the percentage of Malay language usage used by the people of West Kalimantan reached 20.45% (1,615,978 million people) of the total population of West Kalimantan [2]. The efforts to preserve the Pontianak Malay language in order not to become extinct and abandoned because of the influence of globalization must still be done, especially by using text-to-speech technology.

Text-to-speech is a process in which input text is first analyzed, then processed and understood, and then the text is converted to digital audio and the spoken [3]. To develop a speech synthesis to Pontianak Malay Language in order to preserve the local language, predicting pauses from text is an essential part of the text-to-speech system. The presence of pauses supports listeners in parsing the speech stream and enables them to better digest the incoming information [4]. Speech pauses are obtained from beheading phrases. Phrases are grammatical units consisting of one or more words [5]. To get phrases from a sentence can use the chunking method by structuring speech based on grammar rules. Speakers and listeners produce and process language in chunks [21]. In addition to being a component in parsing, chunkers are also used for the development of different natural language processing applications such as information retrieval, information extraction, named entity recognition, etc [22]. The use of chunking helps readers understand the provisional structure of a text and then aids the reader in restructuring and organizing the content of each sentence. The chunking method can use the RegexpParser function in the Natural Language Toolkit to cut sentences into phrases based on the Part of Speech (PoS) type [6]. A regex parser uses a regular expression defined in the form of grammar on top of a POS-tagged string. Grammar rules are needed to define the structure of a chunk. Chunk represents sentence fragments that occur when reading all sentences [7]. Based on this, a pause break can be determined using phrases from the chunking method.

Research on chunking or can be called shallow parsing in Pontianak Malay has been done, where the grammar rules were developed by structuring sentences into S-P-O-K (subject, predicate, object, and adverb) rule [8]. The test results obtained in the form of total f-measure value is 0.64. Recall and precision values for single sentences are 0.78 and 0.74, and compound sentences are 0.67 and 0.57. The ruled that used is only grammar rule and did not check for the pause's type. Of the 168 sentences, the match value with speaker pauses is 40.4% or 68 sentences. The researcher then explained this is because the rule is based on the sentence structure so the phrases did not refer to the pause phrase from the speaker.

Pause is an essential element in the analysis of a text, which also gives good control over interactions during the processes of text reading and explanation of understanding [24]. Insertion of the right amount of pauses at the right places adds to the naturalness of the synthesized speech [9]. Appropriate pausing in the speech can enhance the intelligibility and make the speech more persuasive [18]. Pause also was used to indicate that upcoming words are important and give a sign to the listeners that they should pay attention to those words [19]. There are two factors that influence the speech pausing style, speaker doubts when speaking and breathing method [10]. Abney (1991) explained that when we were reading a sentence, we tend to group words into phrases [7]. Thus, a pause occurs not only based on the influence of the S-P-O-K rules but can be influenced by the speakers themselves.

There is some research about the pause break prediction that has been done which is related to this study. Research about a pause break in English Corpus has been done by using nltk_lite's regular expression chunk parser [11]. There were two tests, one to the input without full stop and comma with 40.5% value, and the other is input with full stop and comma with 43.5%.In this research, nltk_lite's regular expression chunk parser can be used to predict the pause in the English corpus. There is research for the Chinese language based on a maximum entropy model. This used the PoS model and PoS model and lexical to predict phrase break. The result is 62.91% accuracy for PoS model and 65.24% accuracy for PoS and lexical model [12].

In other research, a pause can be predicted by the Hidden Markov model in the Indonesian Language [13]. The research uses the PoS tag tool as one of the features for HMM from Wicaksono's research in 2010 [14]. The result of the recall test is 13.2%, precision with 36.4%, and f-score with 19.4%.

Based on the description above, the researcher intends to develop new grammar rules and pause rules based on the analysis of speaker's pause to categorize chunk phrases in Pontianak Malay language by chunking method to increase the accuracy of pauses prediction in Pontianak Malay sentences so it can be used to develop a good Pontianak Malay Language speech synthesis system. This new PoS tag for Pontianak Malay Language also made in this research.

2. Research Methods

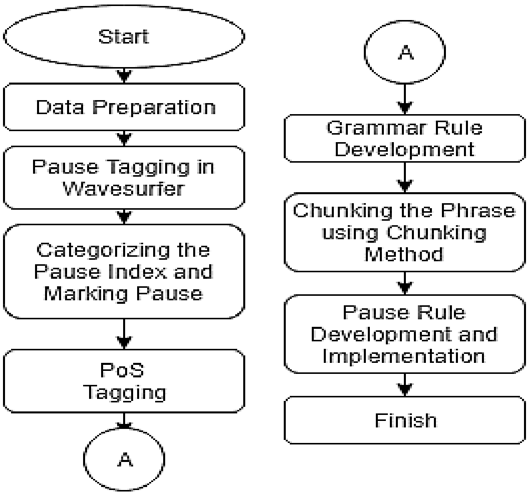

Start
Data Preparation
Pause Tagging in Wavesurfer
Categorizing the Pause Index and Marking Pause
PoS Tagging
A
A
Grammar Rule Development
Chunking the Phrase using Chunking Method
Pause Rule Development and Implementation
Finish

Figure 1. Research Methods

2.1. Data Preparation

The data used is a corpus of 500 Malay Pontianak language sentences from “Sepok”, a Pontianak Malay Language Book [15] consisting of single sentences and compound sentences and each sentence is recorded and spoken by a male speaker who is fluent in the Malay dialect of Pontianak with a daily speaking style. The recordings are stored in a WAV audio format, with 16-bit resolution and 44100 Hz sampling rate.

2.2. Pause Tagging in Wavesurfer

The prepared sound file is then processed using the wavesurfer application to mark the phonemes and pause event. The pause event occurred when the sound wave signal in wavesurfer is flat, which is to identify that the speaker is taking a pause when he is speaking. Each pause event is marked with a “sil” and stored in a file with the format * breaks.

2.3. Categorizing the Pause Index and Marking the Pause in the Pontianak Malay Sentence Text

After all sound files are marked, the “sil” data is analyzed and categorized as a paused index. The sentence then will be marked with a paused index by matching the duration of pause from the sound file that has been marked with “sil”. Table 1 presents the pausing index to determine how long duration for pause “1” and pause “2”.

Table 1. Pause Index

Pause Index	Explanation	Duration of pauses (in second)
0	No pause	0 - < 0.025
1	Short pause	0.023 - <= 0.33
2	Long pause	> 0.33
,	Comma	,
.	End of sentence	.

In table 1, the duration of pauses for pause index “0” is 0 until 0.025 seconds. To mark a paused index for 1 (symbolized as “/1”) is for the duration of sil in the sound file in 0.023 until 0.33 second. For the pause index 2 (symbolized as “/2”) or can be called long paused is for the duration of sil that bigger than 0.33 second. For a comma and full stop, the symbol is the same.

2.4. PoS (Part-of-Speech) Tagging in Pontianak Malay Language Sentences

The 500 Malay Pontianak language is tagged with Pontianak Malay part-of-speech tagger made in this research. Part of Speech Tagging or word class labeling is a process that gives a word class label to each word in sentence or text [20]. PoS Tagging is one of the stages of Natural Language Processing to determine the class of words [23]. Word class consists of adjectives, nouns, verbs, adverbs, prepositions, pronouns, conjunction, etc. This part-of-speech tagger is made for Pontianak Malay Language based on the other PoS set references [8][16][17]. Table 2 presents the Pontianak Malay part-of-speech tag.

Table 2. Part-of-Speech Tag for Pontianak Malay

No	PoS	Description	Example
1	VBR	Reduplication Verb	Jalan-jalan, poto-poto
2	VBK	Conjugation Verb	Bersalam-salam, berputar-putar
3	VBT	Transitive Verb	Makai, nenggek, njajah
4	VBI	Intransitive Verb	Betanyak, balek, nuron
5	IN	Prepostion	di, ke, dari , pade
6	UH	Interjection	Oi, woi, alamak
7	AR	Articulus	Sang, si
8	RP	Particle	pon, lah, jak
9	JJ	Adjective	kaye, lawar, pandai, budoh
10	CON	Conjunction	dan, kalok
11	OP	Open Parenthesis	({ [
12	CP	Close Parenthesis	) }]
13	.	Sentence Terminator	.! ? …
14	.	Comma	,
15	:	Colon	: :
16	SYM	Symbol	*%#&@
17	CR	Currency	Rp, $
18	MD	Modal	nak, haros
19	NEG	Negation	bukan, jangan , tadak
20	SL	Slash	/
21	DS	Dash	-
22	QT	Quotation	" '
23	WP	WH-Pronoun	Ape, siape, berape
24	WDT	WH-Determiner	Ape, siape, barangsiape
25	DT	Determiner	ini, ni , tu, tu, tuh
26	FW	Foreign Word	wonderful, story
27	US	Unit Symbol	Gr, Kg, Cm
28	CDP	Primary Numeral	Satu, duak, tige
29	CDO	Ordinal Numeral	Kesatu, Keduak, ketige
30	CDI	Irregular Numeral	Beberape, segale, semue
31	CDF	Fraction Numeral	Setengah, seperempat
32	CDA	Auxiliary Number	biji, ekor, buah, orang
33	CDC	Collective Numeral	ratusan, ribuan, pulohan
34	RB	Adverb	paleng, sementara
35	WPRB	WH-Adverb	Cemane, ngape
36	FRB	Adverb of Frequency	jarang, sering, kadang-kadang
37	DRB	Adverb of Degree	agak, hamper, cukop
38	TRB	Adverb of Time	udah, belom, dulok, sekarang
39	PRP	Personal Pronoun	aku, saye, kau, die
40	PRL	Locative Pronoun	sanak, sine, situk
41	PRN	Number Pronoun	satu-satunye, dua-duanye
42	NNP	Proper Noun	Eropa, Indonesia, Belanda
43	NNG	Genitive Common Noun	bukunye, rumahnye
44	NNC	Countable Common Noun	buku, rumah, karyawan

No	PoS	Description	Example
45	NNU	Uncountable Common Noun	aek, gula, nasi, ujan
46	NN	Common Noun	Martabat, janji

There is 46 Part-of-Speech tags that made in this research. We can look in table 2, for example for words like “Oi, Woi, Alamak” in table 2 number 6 is categorized as PoS “UH” or Interjection. So, if there is a sentence like “Alamak!”, it will be tagged in PoS became “Alamak/UH ./!”.

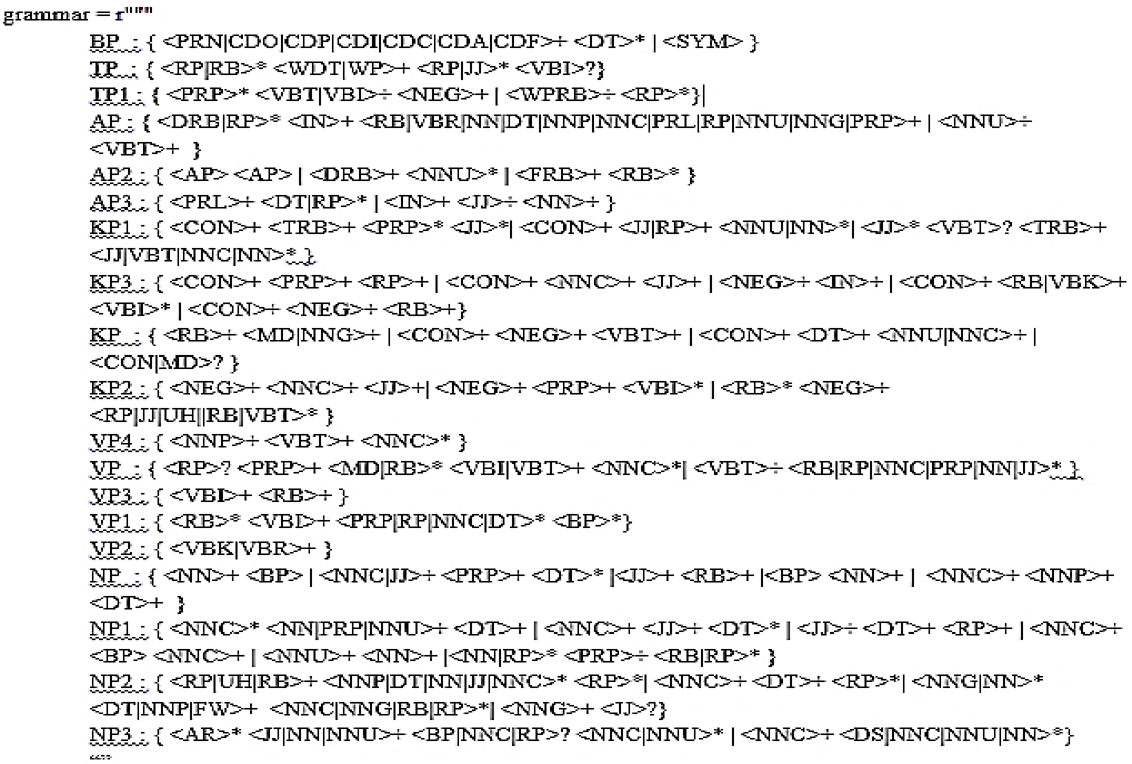
2.5. Grammar Rule Development

Pause event data from point 2.3 and the corpus tagged with PoS from point 2.4 then be analyzed to make grammar rule and pause rule. Grammar rule is for the chunking process. This grammar rule classifies phrases into six types of phrases: TP (Questioning Phrases), BP (Numeric Phrases), NP (Noun Phrases), KP (Connection Phrases), VP (Verb Phrases), and AP (Adverb Phrases). New grammar rule is made by analyzing the pause event from speaker that occurred in the sentences make the pause segment into a chunking phrase rule with the help of regular expression.

Table 3. Regular Expression Characters Meaning

Characters	regular expressions of characters meaning
<>	Determination of part-of-speech tags
?	nothing or one of the previous items
*	Nothing or more than previous items
+	One or more than previous items
|	Matching one item with another

The result of the analysis is 19 new grammar rules for Malay Pontianak language based on the pause event from the native speaker.



```
grammar = r"""
        BP : { <PRN|CDO|CDP|CDI|CDC|CDA|CDF>+ <DT>* | <SYM> }
        TP : { <RP|RB>* <WDT|WP>+ <RP|JJ>* <VBI>?}
        TP1 : { <PRP>* <VBT|VBI>+ <NEG>+ | <WPRB>+ <RP>*}|
        AP : { <DRB|RP>* <IN>+ <RB|VBR|NN|DT|NNP|NNC|PRL|RP|NNU|NNG|PRP>+ | <NNU>+
        <VBT>+  }
        AP2 : { <AP> <AP> | <DRB>+ <NNU>* | <FRB>+ <RB>* }
        AP3 : { <PRL>+ <DT|RP>* | <IN>+ <JJ>+ <NN>+ }
        KP1 : { <CON>+ <TRB>+ <PRP>* <JJ>*| <CON>+ <JJ|RP>+ <NNU|NN>*| <JJ>* <VBT>? <TRB>+
        <JJ|VBT|NNC|NN>* }
        KP3 : { <CON>+ <PRP>+ <RP>+ | <CON>+ <NNC>+ <JJ>+ | <NEG>+ <IN>+ | <CON>+ <RB|VBK>+
        <VBI>* | <CON>+ <NEG>+ <RB>+}
        KP : { <RB>+ <MD|NNG>+ | <CON>+ <NEG>+ <VBT>+ | <CON>+ <DT>+ <NNU|NNC>+ |
        <CON|MD>? }
        KP2 : { <NEG>+ <NNC>+ <JJ>+| <NEG>+ <PRP>+ <VBI>* | <RB>* <NEG>+
        <RP|JJ|UH||RB|VBT>* }
        VP4 : { <NNP>+ <VBT>+ <NNC>* }
        VP : { <RP>? <PRP>+ <MD|RB>* <VBI|VBT>+ <NNC>*| <VBT>+ <RB|RP|NNC|PRP|NN|JJ>* }
        VP3 : { <VBI>+ <RB>+ }
        VP1 : { <RB>* <VBI>+ <PRP|RP|NNC|DT>* <BP>*}
        VP2 : { <VBK|VBR>+ }
        NP : { <NN>+ <BP> | <NNC|JJ>+ <PRP>+ <DT>* |<JJ>+ <RB>+ |<BP> <NN>+ |  <NNC>+ <NNP>+
        <DT>+  }
        NP1 : { <NNC>* <NN|PRP|NNU>+ <DT>+ | <NNC>+ <JJ>+ <DT>* | <JJ>+ <DT>+ <RP>+ | <NNC>+
        <BP> <NNC>+ | <NNU>+ <NN>+ |<NN|RP>* <PRP>+ <RB|RP>* }
        NP2 : { <RP|UH|RB>+ <NNP|DT|NN|JJ|NNC>* <RP>*| <NNC>+ <DT>+ <RP>*| <NNG|NN>*
        <DT|NNP|FW>+  <NNC|NNG|RB|RP>*| <NNG>+ <JJ>?}
        NP3 : { <AR>* <JJ|NN|NNU>+ <BP|NNC|RP>? <NNC|NNU>* | <NNC>+ <DS|NNC|NNU|NN>*}
        """
```

Figure 2. Grammars Rule for Chunking Process

The purpose of the grammar rules in Figure 2 is to be used in the next chunking method.This rule will make the word in sentences that we have been input to be categorized in a phrase that has been made in the rule. For example in rule 3 in figure 2 :

TP1 : { <PRP>* <VBT\VBI>+ <NEG>+ | <WPRB>+ <RP>* } ,

If we have a sentence that consists of word in PoS that included in that rule, the sentence will be cropped into that rule name, for example TP1. For example in sentences “Ikot ndak”, if it tagged with PoS in Table 2 it became “Ikot/VBI ndak/NEG”. When we read the sentences by the grammar rules, the rules would categorize it as TP1 in rule 3 because it contained the same pattern with the rule and became :

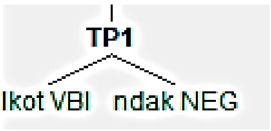


Figure 3. Example of grammar rules

2.6. Chunking the Phrase Using Chunking Method

The chunking process is made to chop sentences into pause phrases using RegexpParser in NLTK. The process of chunking can be seen in Figure 4.

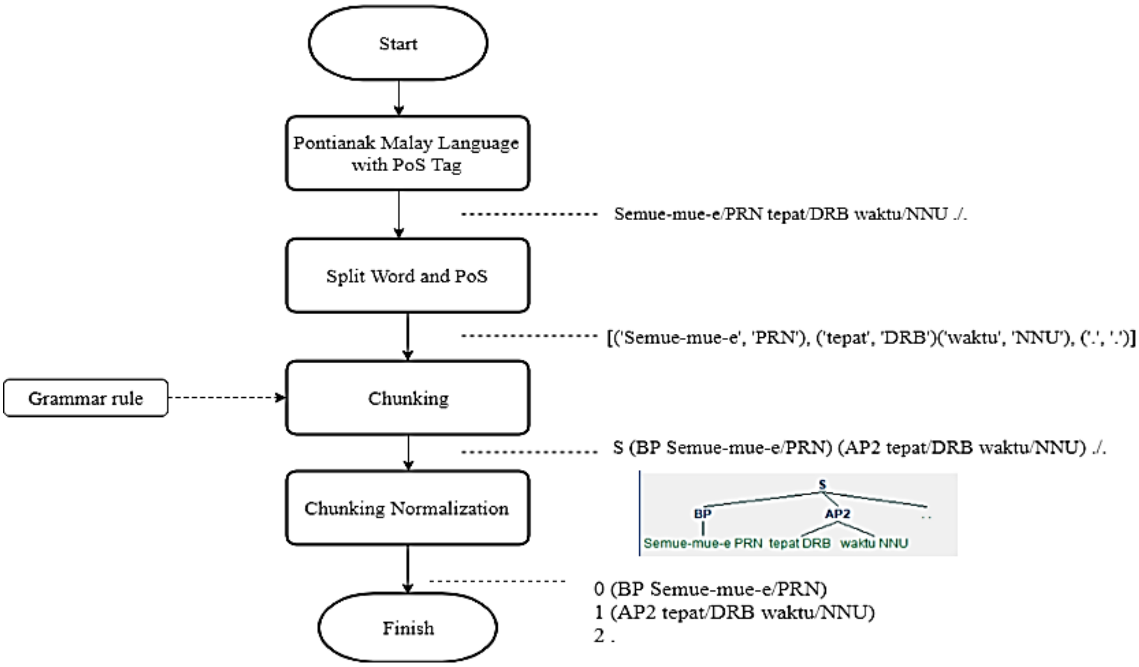

Start
Pontianak Malay Language with PoS Tag
Semue-mue-e/PRN tepat/DRB waktu/NNU ./.
Split Word and PoS
[('Semue-mue-e', 'PRN'), ('tepat', 'DRB')('waktu', 'NNU'), ('.', '.')]
Grammar rule
Chunking
S (BP Semue-mue-e/PRN) (AP2 tepat/DRB waktu/NNU) ./.
Chunking Normalization
S
BP
AP2
Semue-mue-e PRN
tepat DRB
waktu NNU
0 (BP Semue-mue-e/PRN)
1 (AP2 tepat/DRB waktu/NNU)
2 .
Finish

Figure 4. Chunking Process

Using NLTK, when we input the Pontianak Malay Language with PoS Tag, the sentences then will be identified by the PoS label then will be processed by grammar rule to be chunked into chunking phrases. We can look in figure 4, when we have a Pontianak Malay Language Sentence that has been tagged with PoS :

“Semue-mue-e/PRN tepat/DRB waktu/NNU ./.” ,

the next step to do is to split the word and the PoS tag so it can be processed in the next step. After that, the grammar rule in figure 1 will categorize each word into phrases that have been formed in grammar rule. In the example, the sentences are categorized into “ Rule BP : (BP Semue-mue-e/PRN) and Rule AP2 (AP2 tepat/DRB waktu/NNU). All of the sentences in this research are processed in this step so it can be analyzed to get pause rule and can be implemented to make a pause predict.

2.7. Pause Rule Development and the Implementation

Phrase fragments from the chunking process are analyzed to get the type of pause that occurs based on the incidence of the speaker. The pause type consists of two short pauses (symbolized as / 1) and long pauses (symbolized as / 2). The results of the analysis are then used as a pause rule to mark short pauses and long pauses at the pause prediction stage. Figure 5 presents the process of pause rule checking.

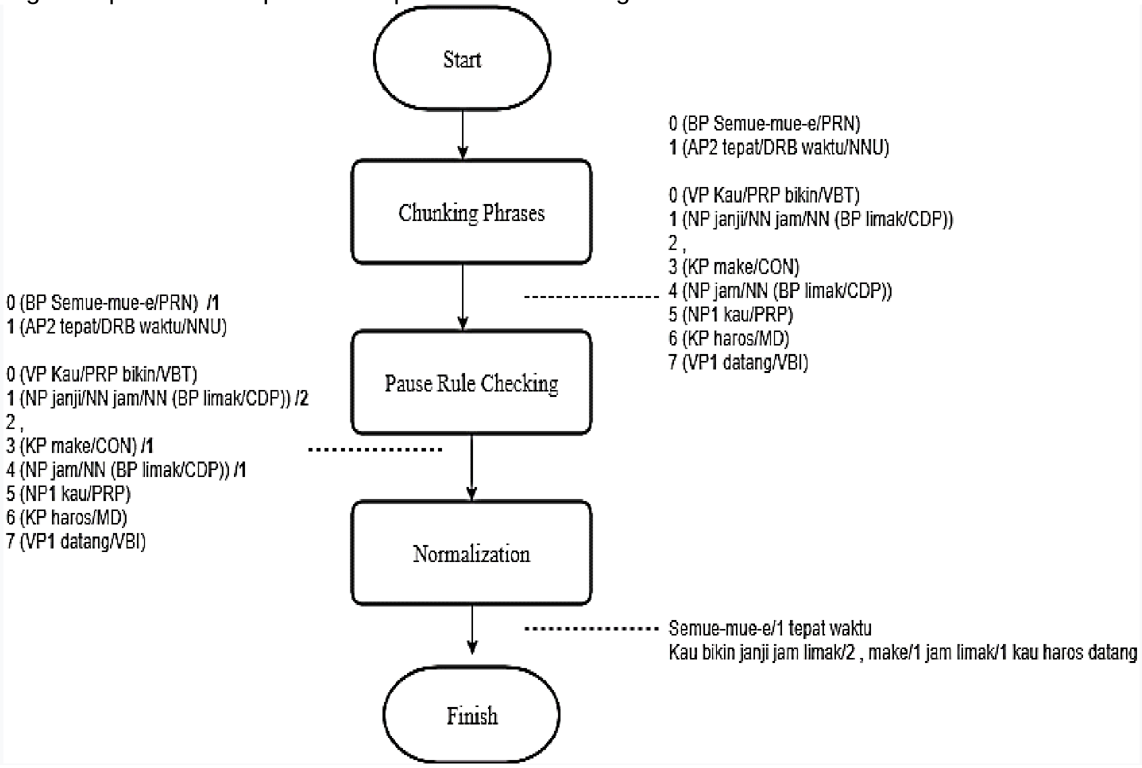

Start
Chunking Phrases
0 (BP Semue-mue-e/PRN)
1 (AP2 tepat/DRB waktu/NNU)
0 (VP Kau/PRP bikin/VBT)
1 (NP janji/NN jam/NN (BP limak/CDP))
2 ,
3 (KP make/CON)
4 (NP jam/NN (BP limak/CDP))
5 (NP1 kau/PRP)
6 (KP haros/MD)
7 (VP1 datang/VBI)
Pause Rule Checking
0 (BP Semue-mue-e/PRN) /1
1 (AP2 tepat/DRB waktu/NNU)
0 (VP Kau/PRP bikin/VBT)
1 (NP janji/NN jam/NN (BP limak/CDP)) /2
2 ,
3 (KP make/CON) /1
4 (NP jam/NN (BP limak/CDP)) /1
5 (NP1 kau/PRP)
6 (KP haros/MD)
7 (VP1 datang/VBI)
Normalization
Semue-mue-e/1 tepat waktu
Kau bikin janji jam limak/2 , make/1 jam limak/1 kau haros datang
Finish

Figure 5. Pause Rule Process

The pause rule that has been made will be implemented in this pause rule process. The phrase fragments from figure 3, will be processed in pause rule checking. For example :

(BP Semue-mue-e/PRN) and **(AP2 tepat/DRB waktu/NNU),** in pause rule when BP run into AP2 then it will be marked as short pause (symbolized as "/1") and became :
(BP Semue-mue-e/PRN)/1 (AP2 tepat/DRB waktu/NNU) and the final sentence would became:
"Semue-mue-e/1 tepat waktu.
For another example, if the phrase fragments are:
(VP Kau/PRP bikin/VBT) (NP janji/NN jam/NN (BP limak/CDP) , (KP make/CON) (NP jam (BP limak/CDP))….., in pause rule when VP run into NP there is no pause , but when NP run into "," it will be marked as long pause (symbolized as "/2"). If KP runs into NP it will be marked as short pause (symbolized as "/1"), so the phrase fragments became:
(VP Kau/PRP bikin/VBT) (NP janji/NN jam/NN (BP limak/CDP)/2 , (KP make/CON)/1 (NP jam (BP limak/CDP)) ….. , and the final sentence would become:
"Kau bikin janji jam limak/2, make/1 jam limak …………………."

After all the process, then the output from this prediction process is tested using pause break accuracy in one sentence and a test using f-measure, recall, and precision parameters. In chunking method, there is no training processing because it based on the rule that has been made. The prediction system is built in a web form and can be accessed on http://203.24.50.138:8027/prediksi_jeda/.

3. Result and Discussion

This research result is tested using two tests, first is pause break compatibility in one sentence testing and the second test is using precision, recall, and f-measure testing.

3.1. Pause Break Compatibility in One Sentence Testing

This testing is done to see the similarity of the occurrence of pauses in the original sentence of the corpus which has been marked the pause event according to the speech of the speaker and the predicted sentence from the chunking process. The total sentences tested were 500 sentences from speaker sentences and 500 sentences as a result of the chunking process. There are two tests carried out, namely testing using the new rule compared to the previous rule from previous research [8].
The example of the test can be seen in Table 4.

Table 4. Example of Pause Break Compatibility in One Sentence Using New

No	Original Pause from Speaker	Chunking Phrase Prediction	Short Pause + Long Pause		Long Pause	
			Same	Not Same	Same	Not Same
1	Kau bikin janji jam limak/2, make/1 jam limak/1 kau haros datang	Kau bikin janji jam limak/2, make/1 jam limak/1 kau haros datang	√	x	√	x
2	Manelah negare kau tuh nak maju/2 kalok tebiat pemerintah-e tak tentu rudu macam itu	Manelah negare kau tuh nak maju/2 kalok tebiat pemerintah-e/1 tak tentu rudu macam itu	x	√	√	x

For the example of the test compared to the previous research can be seen in table 5. In this test, we only see if the phrase fragment of the pause event is same or not because in the previous research, there is no pause index categorization.

Table 5. Example of Pause Break Compatibility in One Sentence Using Previous Rule

No	Original Pause from Speaker	Previous Rule Prediction	Pause	
			Same	Not Same
1	Kau bikin janji jam limak/2, make/1 jam limak/1 kau haros datang	Kau bikin/ janji jam limak/ , make jam limak/ kau haros datang	x	√
2	Manelah negare kau tuh nak maju/2 kalok tebiat pemerintah-e tak tentu rudu macam itu	Manelah negare kau tuh nak/ maju/ kalok tebiat pemerintah-e/ tak tentu rudu macam itu	x	√

In Tables 6 and 7 we could see the testing results. The result is can be seen in the accuracy columns.

Table 6. Pause Break Compatibility in One Sentence Using New Rule

Testing type	Number of Sentences	Accuracy
The appearance of a short pause and long pause	500	33.6% (168 sentences are correct)
The appearance of a long pause	500	72.8% (364 sentences are correct)

Table 7. Pause Break Compatibility in One Sentence Using Previous Rule

Testing type	Number of Sentences	Accuracy
The appearance of pause	500	10.6% (53 sentences are correct)

The accuracy in the table is obtained from the number of sentences that are correctly divided into all of the numbers of sentences. The accuracy told about the chunking phrase accuracy into predicting pause in Pontianak Melayu Sentences. The chunking phrase has a higher accuracy when predicting a sentence with a long pause. But in the sentence that contains a short pause, the accuracy only 33.6% out of 100%.

From the test, we could also see that the accuracy value from the new rule developed in this research is higher than the previous one. In the previous rule, the rule only makes phrases without knowing which is a short and long pause, so there is no test for the appearance of a long pause.

3.2. Precision, Recall, and F-Measure Testing

The evaluation of the prediction is also evaluated in terms of precision, recall, and F-Measure. Precision is the percentage of correct guessed chunks.It is obtained by the total amount of correct chunking phrase and the wrong fragment in the prediction sentences. Meanwhile recall is the percentage of correct chunks were guessed. It is obtained by the total amount of correct chunking phrases and fragments of pauses that were not taken in the original sentence. F-measure is the harmonic mean of precision and recall.

3.2.1. Precision, Recall, and F-Measure Testing to Long and Short Pause.

The testing for a long and short pause in divided into five tests, namely comparing sentences of 100 sentences, 200 sentences, 300 sentences, 400 sentences, and 500 sentences. The test results can be seen in Table 8 and Figure 6.

Table 8. Summary of Testing Value for Long and Short Pause Testing

No	Number of Sentences	Precision	Recall	F-Measure
1	100	0.449	0.475	0.461
2	200	0.448	0.475	0.462
3	300	0.448	0.475	0.461
4	400	0.448	0.475	0.462
5	500	0.448	0.475	0.462

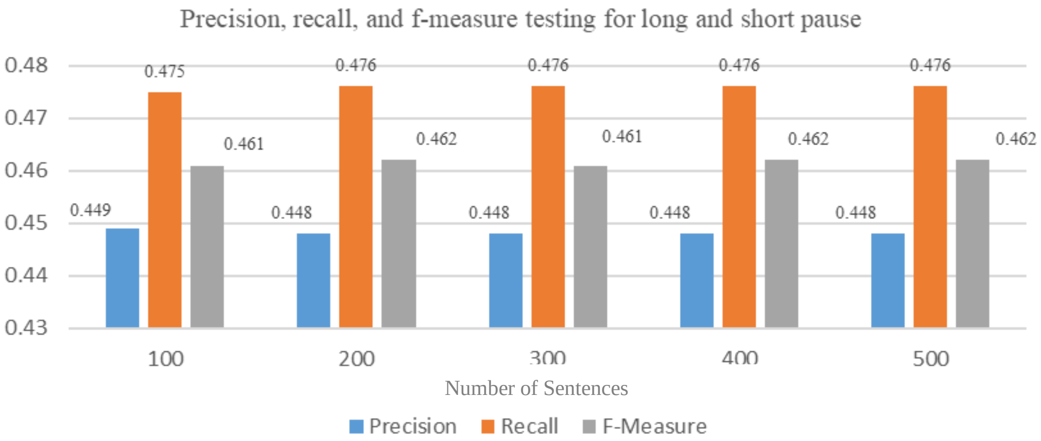


Figure 6. Precision, Recall, and F-measure Testing Value Chart for the Long and Short Pause

In this testing, we could see in Table 8 and Figure 6, the precision value or the percentage of correct guessed chunks for the sentences is almost the same and the recall is same. The harmonic mean or the f-measure value is almost the same in 0.46. The value that almost same showed in Figure 6 is meant that the chunking phrase makes in this research based on the rule to predict the pause predict is consistent in predicting the pausing index.

The value in the test which is in the range of 0.4 due to chunking prediction is not accurate due to the low precision value. Many irrelevant phrases or pause phrases that have not been properly formed. This wrong pause phrase is because the grammar rule forms phrases according to the type of post that appears in the sentence. Short pauses have a pattern of pauses that vary from the speaker which causes the appearance of pauses to be unequal.

3.2.2. Precision, Recall, and F-Measure Testing to Long Pause

The testing for a long pause in divided into five tests, namely comparing sentences of 100 sentences, 200 sentences, 300 sentences, 400 sentences, and 500 sentences. The test results can be seen in Table 9 and Figure 7.

Table 9. Summary of Testing Value for Long Pause Testing

No	Number of Sentences	Precision	Recall	F-Measure
1	100	0.746	0.703	0.724
2	200	0.746	0.702	0.723
3	300	0.746	0.702	0.723
4	400	0.746	0.701	0.723
5	500	0.746	0.701	0.723

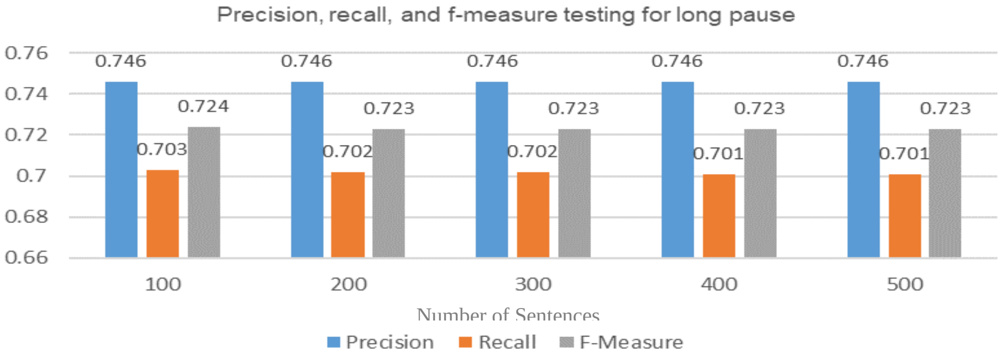


Figure 7. Precision, Recall, and F-measure Testing Value Chart for the Long Pause

In this testing, we could see in Table 9 and Figure 7, the precision value or the percentage of correct guessed chunks for the sentences and the recall is same. The harmonic mean or the f-measure value is almost the same in 0.72. The value that almost same showed in Figure 6 is meant that the chunking phrase makes in this research based on the rule to predict the pause predict is consistent in predicting the pausing index.

The precision testing value shows the same number at 0.746 which means that the rule grammars and the pause rule succeed in predicting the right fragment for all sentences from 100 sentences to 500 sentences. The recall value is more varied because there are still fragments of phrases that do not match the speakers' pause phrases because the rules do not match. The f-measure value has almost the same value and is classified as good which is 0.72. The prediction of long pauses has better and higher values because based on the speakers' pauses, the location of the long pauses tends to have a stop pattern in the same phrase so that the grammar rule and the paused rule created can predict the gap well.

3.3. Analysis of The Test Results

Based on the results of pause break compatibility in one sentence, the value of the accuracy has increased by 23% value. The new rule has better accuracy in predicting pause based on the speaker's speech.

In precision, recall, and f-measure testing, based on Tables 8 and 9, long pause prediction has a better value. This is because, based on analysis while making grammar and pause rule, the long pause is easier to be formed than the short pause. Based on the speaker, the short pause has a varied and different pattern in each sentence which makes the rule cannot predict all the testing sentence into a perfect prediction. This is also due to the imperfect labeling word class that make rule cannot cut phrases into accurate prediction according to the speaker's phrase.

Table 10. Pause Comparison

No	Pause from Speaker	Pause from System
1	kame/PRP ni/DT **jaim/VBI/1 tang/IN** atas/NN kapal/NNC	kame/PRP ni/DT **jaim/VBI/1 tang/IN** atas/NN kapal/NNC
2	**Naekan/VBT ke/IN** atas/NN kapal/NNC klotok/NNC.	**Naekan/VBT/1 ke/IN** atas/NN kapal/NNC klotok/NNC.

In Table 10, we can see the difference in the speaker's pause and system. In the first sentence, after the word with pos label verb VBI, a short pause occurs before the preposition "tang" with the label "IN". This is because the rule is set to have a short pause before "IN" for a word like "tang".The grammar and pause rules predict the same results as the speakers. Meanwhile, in the second sentence, verb VBT and IN do not pause. Because the "IN" PoS is assigned to a word named "ke". So the prediction results are not accurate.

4. Result and Discussion

Based on the test results, the new grammar rule and pause rule that formed a chunking phrase can predict the pause in Pontianak Malay language with accuracy about 33.6% for short pause and long pause in one sentence, and 72.8% for the long pause. This value has a better number than the previous rule. The best value is for long pause with 72.8% compatibility with speaker's pause and precision value with 0.74, recall with 0.70 and f-measure with 0.723. The chunking phrase can be implemented to develop a text-to-speech system for Pontianak Malay Language.

References

[1] M. Dwi Etsa Putra, "Pengaruh Metode Dictionary Lookup Pada Proses Cleaning Korpus Terhadap Akurasi Mesin Penerjemah Statistik Bahasa Indonesia-Bahasa Melayu

Pontianak," Universitas Tanjungpura, 2018.

[2] N. dan S. H. Akhsan, *Hasil Sensus Penduduk 2010: Kewarganegaraan, Suku Bangsa, Agama dan Bahasa Sehari-hari Penduduk Indonesia*. Jakarta: Badan Pusat Statistik, 2010.

[3] A. Trivedi, N. Pant, P. Shah, S. Sonik, and S. Agrawal, "Speech to text and text to speech recognition systems-A review," *IOSR Journal of Computer Engineering*, vol. 20, no. 2, p. 39, 2018.

[4] N. Braunschweiler and R. Maia, "Pause prediction from text for speech synthesis with user-definable pause insertion likelihood threshold," in *INTERSPEECH 2016*, 2016, p. 3191.

[5] A. Wahab Syahroni, J. Santoso, and E. Setyati, "Pendekatan Rule Handmade untuk Menentukan Klausa Bahasa Indonesia," in *E-Proceedings KNS&I STIKOM Bali 2017*, 2017, pp. 598–603.

[6] R. J. Prathibba and M. C. Padma, "Shallow Parser for Kannada Sentences Using Machine Learning Approach," *International Journal of Computational Linguistics Research Vol. 8 Number 4*, pp. 158–170, 2017.

[7] S. Abney, "Parsing By Chunks. In Berwick, Abney, and Tenny (eds)," 1991.

[8] M. I. Kamiludin, "Prediksi Jeda Pada Ucapan Bahasa Melayu Pontianak dengan Menggunakan Metode Shallow Parsing," Universitas Tanjungpura, 2017.

[9] P. Arulmozhi and A. G. Ramakrishnan, "Prediction of Pauses in TTS - Tamil," in *Conference: Tamil internet 2010*, 2010.

[10] S. Darjdowidjojo, *Psikolinguistik, Pengantar Pemahaman Bahasa Manusia*. Jakarta: Yayasan Obor Indonesia, 2005.

[11] C. Brierley and E. Atwell, "Corpus-Based Evaluation of Prosodic Phrase Break Prediction Using nltk_lite;s Chunk Parser to Detect Prosodic Phrase Boundaries in the Aix-MARSEC Corpus of Spoken English," United Kingdom, 2007.

[12] L. Jian-feng, H. Guo-ping, Z. Wan-ping, and W. Ren-hua, "Chinese Prosody Phrase Break Prediction Based on Maximum Entropy Model," in *INTERSPEECH 2004*, 2004.

[13] A. Teguh Nugraha, "Prediksi Jeda Dalam Ucapan Kalimat Bahasa Indonesia Dengan Hidden Markov Model," Universitas Tanjungpura, 2014.

[14] A. F. Wicaksono and A. Purwarianti, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia," in *Conference: 4th International MALINDO (Malaysian-Indonesian Language) Workshop*, 2010.

[15] P. J. Sujarwo, *Sepok: Cerite Orang Kampong, yang Kampongan, di Kampong Orang*. Pontianak: Pijar Publishing, 2010.

[16] E. Rahayu Setyaningsih, "Part of Speech Tagger Untuk Bahasa Indonesia Dengan Menggunakan Modifikasi Brill," *Dinamika Teknologi*, vol. 9, pp. 37–42, 2017.

[17] M. Adriani and H. Riza, "Research Report on Local Language Computing: Development of Indonesia Language Resources and Translation System," 2009.

[18] P.Sarkar and K.Sreenivasa Rao, "Data-Driven Pause Prediction for Synthesis of Storytelling Style Speech Based On Discourse Modes," In: 2015 IEEE International Conference on Electronics, Computing and Communication Technologies, 2015.

[19] Q. Truong Do, S.Sakti,G.Neubig, T.Toda and S.Nakamura, "Improving Translation of Emphasis with Pause Prediction in Speech-to-Speech Translation Systems," Japan: Nara Institute of Science and Technology, 2015.

[20] R.Manurung, "Tutorial: Pengenalan Terhadap POS Tagging dan Probalistic Parsing," Workshop Nasional INACL, 2016.

[21] R.Niu and T.Osborne, "Chunks are Components: A Dependency Grammar Approach to The Syntactic Structure of Mandarin," Lingua: Elsevier, 2019

[22] A. Ibrahim and Y.Assabie, "Amharic Sentence Parsing Using Base Phrase Chunking,", In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing, CICLing 2014.

[23] A. Subhan Yazid and A.Fatwanto, "Penentuan Kelas Kata Pada Part of Speech Tagging Kata Ambigu Bahasa Indonesia," Jurnal Informatika Sunan Kalijaga, vol.2, No.3, pp. 157-166, 2018

[24] S. Denisleam-Molomer, S.Trausan-Matu, P.Dessus, and M.Bianco," Analyzing Students Pauses During Reading and Explaining A Story," RoEduNet International Conference: Networking in Education and Research 2015, Craiova, Romania, pp.90-93, 2015