Improving NORMALS Using Modified Baudot-Murray Code

Andreyadi Wibowo
Informatics Engineering Graduate School of Telkom University
Jl. Telekomunikasi Terusan Buah Batu
Bandung 40257, Indonesia
+62 22 7564 108
andreyadi@students.telkomuniversity.ac.id

Ari Moesriami Barmawi
Informatics Engineering Graduate School of Telkom University
Jl. Telekomunikasi Terusan Buah Batu
Bandung 40257, Indonesia
+62 22 7564 108
mbarmawi@melsa.net.id

ABSTRACT
NORMALS or Normal Linguistic Methodology Steganography is a steganography method based on noiseless steganography paradigm or Nostega. In this method, a message is embedded into cover text by modifying the external input of a Natural Language Generation (NLG) system that produces text. The main problem of NORMALS method is small embedding capacity. To solve this problem, this research proposed some method to improve NORMALS method. A better embedding capacity can be achieved by modifying the character encoding used in this research. In addition to modifying the character encoding to make it more efficient, this research also ensures that all the code are evenly distributed, so that in writing the secret message all the code in modified character encoding has almost the same probability to be used. This can reduce suspicion because there is no code that excessively used. The results of the experiments showed that the proposed method has better efficiency in writing the secret message compared to NORMALS, especially for secret messages in the same language with a corpus that is used to modify the character encoding.

CCS Concepts
• **Security and privacy → Authorization.**

Keywords
Steganography; Character Encoding; Text Generation.

1. INTRODUCTION
Recent advances in communications technology in the past years has provided many benefits but it also makes the digital data exchange more susceptible to eavesdropping and malicious intervention. This makes the issues of security becomes more relevant today than ever. Based on this reason, researchers did their best to find different ways for maintaining the confidentiality of the information which prevents the unauthorized parties to access them.

One of the methods for maintaining the information confidentiality stenography is to conceal the information within another media like text, audio or video in such a way that no one knows or realizes that there is a secret message beside the sender and recipient.

ICCNS 2016, November 26-29, 2016, Singapore, Singapore

DOI: http://dx.doi.org/10.1145/3017971.3017984

In 2009, Desoky introduced noiseless Steganography or Nostega paradigm, which he described as a paradigm for designing systems which make the existence of data hidden in the cover as natural as possible so that no shortage of linguistic or strangeness was introduced as a side effect [1]. He proposed some methods like Grapstega [2], Chestega [3], and Edustega [4].

Another research based on Nostega is NORMALS (Normal Linguistic Steganography). NORMALS employs Natural Language Generation (NLG) techniques to create a noiseless and valid text cover by modifying the non-random series input of an NLG system in order to embed the secret message in the generated text [5].

This paper proposed a steganography method which utilizes the use of a modified character encoding to improve the embedding capacity of NORMALS.

2. RELATED STUDIES
NORMALS is a noiseless steganography method which hides the secret message in a cover text by modifying the input of an NLG system which generated intended text.

NORMALS architecture consists of two main modules as shown in Fig. 1 [5].

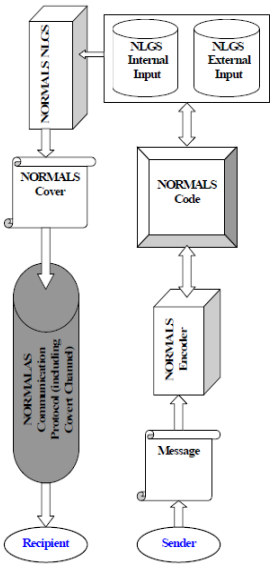

Figure 1. NORMALS Architecture [5].

The function of each main block in NORMALS method can be described as follow:

1) NORMALS Encoder: in this module, the secret message is converted from text (8-bit ASCII) to binary to create NORMALS Code.

2) NORMALS NLG: The binary that contain in NORMALS Code is used as an external input for NORMALS NLG to create a NORMALS Cover (the cover text).

Because NORMALS method uses ASCII to encode the secret message and this research uses questionnaire with five possible answers for NLG, they always need 4 questions to embed 1 character in the secret message. For example:

For questionnaire with five possible answers, if A = 00, B = 01, C = 10, and D = 11, then secret message "**MAKAN MALAM**" (dinner) will be encoded using ASCII into:

01001101_2 01000001_2 01001011_2 01000001_2 01001110_2
00100000_2 01001101_2 01000001_2 01001100_2 01000001_2
01001101_2

The answers that we have to fill into NORMALS NLG are:

M	A	K	A	N	
BADB	BAAB	BACD	BAAB	BADC	ACAA

M	A	L	A	M
BADB	BAAB	BADA	BAAB	BADB

For detail step by step process how the NORMALS can be used to hide the secret message can be found at "**NORMALS: Normal Linguistic Steganography Methodology**" [2].

3. IMPROVING NORMALS USING MODIFIED BAUDOT-MURRAY CODE

The proposed method utilizes a more efficient character encoding method which has been conformed to a questionnaire which will be used as input for the NLG systems as shown in Fig. 2.

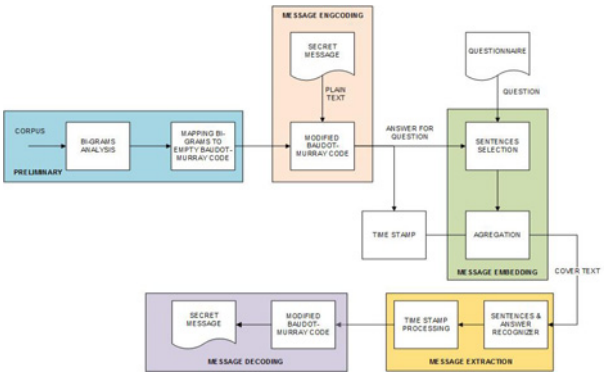

Figure 2. Proposed Method Block Diagram.

3.1 Preliminary Research

For implementing the proposed method, the preliminary research was conducted to choose a character encoding.

To increase the efficiency of character encoding in the improved NORMALS method, it was necessary to seek the character encoding which uses less than 8-bit but still covers all the needed character for writing the secret message. Due to length limitation of the secret message that can be embedded into the cover text, in this research assumed that the character to be used only in the format of alphabets and space, then it need at least 27 symbols or 5-bit character encoding

The choice fell on Baudot-Murray Code or International Telegraph Alphabet No. 2 (ITA2) which use exactly 5-bit for character encoding. On assumption that the questionnaire which will be used use is a commonly found questionnaire with five possible answers, to utilize all the possible answers this research decided to use the encoding in quinary symbols instead binary.

Quinary (base-5) is a numeral system, a writing system for expressing numbers, with five as the base. Therefore quinary can be stated from 0-4, and can be written as 0_5, 1_5, 2_5, 3_5, and 4_5. There are other commonly used numeral system like binary (base-2), decimal (base-10), and hexadecimal (base-16).

To further increase the efficiency the research change the character encoding into n-gram encoding [6], but as a precaution, if the n-gram cannot encode the secret message, the single character encoding is still included in modified character encoding. For this decision, is necessary to increase the size of character encoding from 5^2 to 5^3. As Baudot-Murray code has two sub-sets, "Letter Shift" and "Figure Shift", after mapping the alphabets and space there are still 221 unused codes which can be used to map the n-gram.

N-gram was obtained from analyzing the corpus, collected from **magazine articles in Indonesia language**. From the analyzed corpus, the top 221 bi-grams consist of 96.03% of all bigrams while the top 221 tri-grams only consist of 50.69% of all trigrams. Because the higher percentage means the system will less likely use single character encoding, then it was decided to use bi-grams encoding instead of tri-grams encoding.

The bi-grams mapped into the empty code in the modified Baudot-Murray code by equalizing the probability of appearance of each symbol in the quinary. After all the bi-grams entered into the modified Baudot-Murray table, the probability of appearance of each symbol is as shown in Table 1.

Table 1. The probability of appearance of each symbol

0_5	1_5	2_5	3_5	4_5
58.39%	58.26%	58.14%	57.93%	55.37%

The symbol 4_5 has the lowest probability because the single character alphabets and space was placed at the end of the table where a lot of symbol 4_5 located.

The final modified Baudot-Murray table which was derived from analyzed corpus can be found in Table 2.

3.2 Message Encoding

Message Encoding is a process to encode plain text into quinary symbols by splitting the secret message into bi-grams and find their counterpart in the modified Baudot-Murray Code.

Because each symbol in modified Baudot-Murray code represents 3 questions in the questionnaire, the number of questions needed can be calculated as:

Number of Questions Needed = (Bi-grams + Single Characters + Shift) * 3

Table 2. The Modified Baudot-Murray Code

Symbols	Letter Shift	Figure Shift	Symbols	Letter Shift	Figure Shift
000	A N	[SPACE] G	223	E [SPACE]	I P
001	E S	L T	224	N I	A U
002	R [SPACE]	K T	230	[SPACE] O	R G
003	A S	A B	231	K O	R L
004	D E	C O	232	A D	U D
010	A P	E A	233	E N	[SPACE] W
011	I N	P [SPACE]	234	R I	A A
012	M U	S P	240	T R	L K
013	E P	N K	241	[SPACE] Y	R N
014	I K	I M	242	[SPACE] I	[SPACE] E
020	[SPACE] A	I O	243	N E	I H
021	U R	F A	244	[SPACE] D	G U
022	T E	P O	300	D A	R M
023	E D	Y [SPACE]	301	K E	O T
024	B I	T O	302	L I	O D
030	A T	A W	303	B E	E G
031	I L	H U	304	I T	P U
032	B U	G R	310	[SPACE] J	E V
033	E M	D O	311	M A	I F
034	A M	F I	312	S T	A C
040	O N	E C	313	T [SPACE]	[SPACE] F
041	L E	V I	314	J U	L O
042	M B	K S	320	G G	F [SPACE]
043	G I	O G	321	A Y	P T
044	M E	R S	322	N T	D [SPACE]
100	B A	O [SPACE]	323	[SPACE] S	U H
101	S A	D U	324	U A	M P
102	C A	C U	330	[SPACE] L	S O
103	S U	E H	331	H [SPACE]	F O
104	U [SPACE]	P R	332	[SPACE] T	N C
110	I A	N N	333	N G	G K
111	N [SPACE]	I R	334	[SPACE] K	U M
112	K [SPACE]	U G	340	N D	V E
113	N Y	P I	341	R O	U L
114	P E	K N	342	U S	A A
120	[SPACE] C	Y E	343	A L	B B
121	Y A	N U	344	I [SPACE]	C C
122	[SPACE] P	R B	400	A R	D D
123	E B	N F	401	U T	E E
124	H A	G E	402	O R	F F
130	E K	R P	403	L U	G G
131	S [SPACE]	I B	404	[SPACE] B	H H
132	[SPACE] U	A J	410	A G	I I
133	D I	N J	411	L A	J J
134	W A	I D	412	R U	K K
140	[SPACE] H	U P	413	L [SPACE]	L L
141	S E	I C	414	T U	M M
142	E T	H I	420	K I	N N
143	R E	M O	421	U K	O O
144	A H	R K	422	G [SPACE]	P P
200	U N	N S	423	A I	Q Q
201	K U	O K	424	K A	R R
202	R A	P L	430	O L	S S
203	M I	O P	431	R T	T T
204	I S	N O	432	J A	U U
210	[SPACE] N	[SPACE] V	433	G A	V V
211	[SPACE] M	O S	434	E R	W W
212	T A	I G	440	E L	X X
213	M [SPACE]	W I	441	T I	Y Y
214	O M	C E	442	P A	Z Z
220	N A	S N	443	S I	[SPACE]
221	A K	O B	444	Shift to figures	Shift to letters
222	A [SPACE]	[SPACE] R			

Based on those, the maximum embedding capacity can be calculated as:

$$Maximum\ embedding\ capacity\ (bi-grams) = \left\lfloor \frac{Number\ of\ questions}{3} \right\rfloor$$

or

$$Maximum\ embedding\ capacity\ (characters) = \left\lfloor \frac{Number\ of\ questions}{3} \right\rfloor \times 2$$

Whereas the worst case scenario can be found if the encoder having to switch back and forth between "Figure Shift" and "Letter Shift", then minimum embedding capacity can be calculated as:

$$Min.embedding\ capacity = \left\lfloor \frac{(Number\ of\ questions)}{15} \right\rfloor \times 4 + \left\lfloor \frac{(Number\ of\ questions)\ Mod\ 15}{6} \right\rfloor \times 1$$

3.3 Message Embedding

In message embedding (see Fig. 3) selected questionnaire will be used as the input of the NLG system. This research will use the questionnaire concerning e-money for input of NLG system to generate response letter to each respondent [7]. The questionnaire is a 19 questions e-money questionnaire with five possible answers [6] shown in Table 3.

Based on the answer to each question, the NLG system will select sentences from a template database. Because the NLG system will need all 19 answer to generated a cover text, if the secret text cannot provide it, then the system will supply the rest through a random generator.

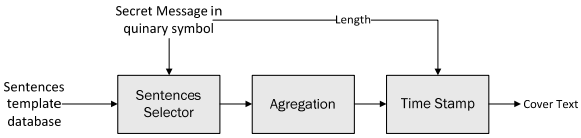

Figure 3. Embedding Process

After the system selects the sentences then it will aggregate the sentences if needed to create more natural sentences. Taking into account that the answers that have been modified to the needs of hiding the secret message could cause conflicting answer between questions, the only relationship between sentences used in this research is simply SEQUENCE or CONTRAST.

The end of the generated text the system will add a time stamp to record the length of the secret message. In this research, it is assumed that the addition of a time stamp at the end of the text is a common disclaimer displayed in a report like this.

Table 3 E-money Questionnaire [6]

A.	Performance Expectancy
PE1	Layanan e-money membantu saya mengelola aktivitas keuangan saya sehari-hari. E-money service helped me to manage my daily financial activity
PE2	Menggunakan layanan e-money membantu saya menyelesaikan transaksi dengan lebih cepat. Using e-money service helped me to complete my transaction more quickly
PE3	Menggunakan layanan e-money membuat saya bisa membayar tepat sesuai nilai transaksi. Using e-money service helped me to complete my transaction more accurately
PE4	Menggunakan layanan e-money membuat saya tidak perlu membawa banyak uang tunai. By using e-money, I do not have to carry a lot of cash.
PE5	Menggunakan layanan e-money membuat saya mendapatkan manfaat khusus di merchant yang bekerja sama dengan penerbit e-money. By using e-money, I can get the special offer from the merhant who have cooperated with e-money provider

C.	Social Influence
SI1	Keluarga saya berpikir bahwa saya seharusnya menggunakan layanan e-money. My family think that I should to use e-money service
SI2	Teman-teman saya berpikir bahwa saya seharusnya menggunakan layanan e-money. My friends think that I should to use e-money service
SI3	Orang-orang di lingkungan sekitar saya (kantor, sekolah, kampus) mendukung saya untuk menggunakan layanan e-money. People in my neighborhood (offices, schools, colleges) supported me to use e-money services.
SI4	Anggota komunitas yang saya ikuti banyak yang menggunakan layanan e-money. Many members of the community which I follow used e-money services.

F.	Price Value
PV1	Saya tidak keberatan dengan biaya awal untuk mendapatkan layanan e-money. I have no objection to initial cost to obtain e-money service
PV2	Saya tidak keberatan dengan biaya transaksi layanan e-money. I have no objection to the transaction fee of e-money service
PV3	Menurut saya, biaya layanan e-money masih wajar. In my opinion, the cost of e-money services is still reasonable.
PV4	Menurut saya, biaya layanan e-money sesuai dengan manfaat yang saya peroleh. In my opinion, the cost of e-money services still in accordance with the benefits I have obtained
PV5	Menurut saya dengan besaran biaya saat ini, layanan e-money memberikan nilai yang baik. In my opinion, compared with the current fee, e-money services provide good value.

I.	Use Behaviour
	Silakan pilih seberapa sering Anda menggunakan layanan e-money untuk item berikut: Please choose how often you use e-money services for the following items:
UB1	Berbelanja Shopping
UB2	Membeli pulsa telekomunikasi atau listrik prabayar Top-up for phone or electricity service
UB3	Membayar tagihan Pay bills
UB4	Transfer Uang Transferring money
UB5	Tarik Uang Tunai Cash withdrawal

3.4 Message Extraction

For message extraction, this research uses two keywords to recognize: one for recognizing the question and one for recognizing the answer for respective questions (see Fig. 4)

Figure 4. Extraction Process

After all the answer for each question has been obtained, the next process is to determine the length of the real message by calculating the time stamp in the secret message.

3.5 Message Decoding

The last section, message decoding, is a process to decode the secret message in quinary symbols which were obtained from the previous section into readable text by using modified Baudot-Murray code.

4. EXPERIMENT RESULT & DISCUSSION

Several scenarios of experiments were conducted to compare NORMALS performance with the proposed method. One of the results of the experiments can be found in Table 4.

Table 4 Sample of Experiments

SECRET MESSAGE	MAKAN MALAM, words with 11 character
ENCODING	343_5 4245 111_5 343_5 411_5 004_5
QUESTIONS NEEDED	18 Questions
COVER TEXT	Anda merasa terbantu dalam mengelola aktivitas keuangan Anda setelah menggunakan e-money. Anda selalu dapat bertransaksi lebih cepat dengan e-money, dan Anda dapat bertransaksi lebih tepat dengannya. Anda tidak pernah sekalipun membawa banyak uang tunai sejak menggunakan e-money. Anda terkadang mendapatkan manfaat khusus dari merchant sejak menggunakan e-money. Keluarga Anda sangat menyarankan Anda untuk menggunakan e-money, walaupun Teman-teman Anda menentangnya. Orang-orang di lingkungan Anda menentang Anda menggunakan e-money. Anda mengikuti komunitas dimana hanya sedikit anggota komunitas yang rutin menggunakan e-money. Anda menilai besaran biaya layanan e-money masuk dalam katagori wajar. Anda setuju dengan besaran biaya awal dari layanan e-money, bahkan Anda sangat setuju dengan besaran biaya transaksi dari layanannya. Anda berpendapat layanan e-money memberikan manfaat yang sangat baik dilihat dari besaran biaya yang diperlukan. Anda berpendapat layanan e-money memberikan nilai yang buruk dilihat dari besaran biaya yang diperlukan. Anda jarang menggunakan layanan e-money untuk berbelanja, bahkan Anda jarang sekali menggunakan layanannya untuk membeli pulsa telekomunikasi atau listrik prabayar. Anda jarang sekali menggunakan layanan e-money untuk membayar tagihan. Anda sering sekali menggunakan layanan e-money untuk transfer uang dan untuk tarik uang tunai. Laporan ini dihasilkan secara otomatis oleh sistem pada tanggal 2016-08-24 jam 13:56:51.79
	You feel that you can manage your financial activities better after using e-money. You can always finish the transaction more quickly with e-money, and also you can do it more precisely. You never carry a lot of of cash since using e-money. Sometimes you get a special benefit of merchants by using e-money. Your family strongly recommend you to use e-money, even if your friends against it. People in your neighborhood oppose the use of e-money. You're following the community where only a few members of the community who regularly uses e-money. You assess the amount of e-money service fees included in the fair category. You agree with the initial fee needed for using e-money, even you completely agree with the the transaction fee of its services. You think of e-money services provide excellent benefits seen from the fees required. You think of e-money services provide poor value seen from the fees required. You seldom use e-money for shopping, altough you rarely use it to top-up the telecommunications or electricity utilities. You rarely use e-money to pay the bills. You often use e-money services to transfer money and to take cash. This report is generated automatically by the system at 2016-08-24 13:56:51.79.

Because of the length of cover text generated by the system varies depending on the answers provided, to make a comparison between methods, this research only comparing saving capacity:

$$\textit{Saving capacity} = \textit{remaining questions using improved NORMALS} - \textit{remaining questions using NORMALS}$$

4.1 Indonesian Language Secret Message

On the ground that the corpus used to create modified Baudot-Murray Code table in Indonesian language, the first experiment naturally use Indonesian language as the secret message.

Based on resume shown in Table 5, the experiments show that the increased efficiency when using Indonesian language as a secret message was varied between secret messages.

Table 5 Experiments Using Indonesian Message

SECRET MESSAGE	PREVIOUS METHOD		PROPOSED METHOD		SAVING CAPACITY
	PREVIOUS METHOD	QUESTIONS NEEDED / REMAINING QUESTIONS	PROPOSED METHOD	QUESTIONS NEEDED / REMAINING QUESTIONS	
DIA (HIS/HER)	$01000100_2\ 01001001_2\ 01000001_2$	12 questions / 7 questions	$332_5\ 222_5$	6 questions / 13 questions	6 questions
LARI (RUN)	$01001100_2\ 01000001_2\ 01010010_2\ 01001001_2$	16 questions / 3 questions	$411_5\ 234_5$	6 questions / 13 questions	10 questions
AWET (DURABLE)	$01000001_2\ 01010111_2\ 01000101_2\ 01010100_2$	16 questions / 3 questions	$444_5\ 110_5\ 444_5\ 134_5$	12 questions / 7 questions	4 questions
PENYU (TURTLE)	$01010000_2\ 01000101_2\ 01001110_2\ 01011001_2\ 01010101_2$	20 questions / -1 questions	$200_5\ 340_5\ 104_5$	9 questions / 10 questions	11 questions
AZIMAT (AMULET)	$01000001_2\ 01011010_2\ 01001001_2\ 01001101_2\ 01000001_2\ 01010100_2$	24 questions / -5 questions	$444_5\ 342_5\ 442_5\ 024_5\ 444_5\ 440_5$	18 questions / 1 questions	6 questions
JABATAN (POSITION)	$01001010_2\ 01000001_2\ 01000010_2\ 01000001_2\ 01010100_2\ 01000001_2\ 01001110_2$	28 questions / -9 questions	$432_5\ 100_5\ 212_5\ 111_5$	12 questions / 7 questions	16 questions
DI KANTOR (IN THE OFFICE)	$01000100_2\ 01001001_2\ 00100000_2\ 01001011_2\ 01000001_2\ 01001110_2\ 01010100_2\ 01001111_2\ 01010010_2$	36 questions / -17 questions	$332_5\ 334_5\ 000_5\ 444_5\ 241_5\ 424_5$	18 questions / 1 questions	18 questions
MANDI PAGI (MORNING BATH)	$01001101_2\ 01000001_2\ 01001110_2\ 01000100_2\ 01001001_2\ 00100000_2\ 01010000_2\ 01000001_2\ 01000111_2\ 01001001_2$	40 questions / -21 questions	$343_5\ 123_5\ 344_5\ 022_5\ 402_5$	15 questions / 4 questions	25 questions

The highest increase obtains when the system use "**LARI**" (run) and "**MANDI PAGI**" (morning bath) as a secret message. "**LARI**" could be separate into bi-grams "**LA**" and "**RI**". Both bi-grams found on the "**Letter Shift**" part, so they did not need to use "**Shift to Figure**" and gave some of the best results. The same result was obtained when encoding "**MANDI PAGI**", bi-grams "**MA**", "**ND**", "**I** ", "**PA**", and "**GI**" can all be found in the "**Letter Shift**" part.

The lowest increase in efficiency, found on "**AWET**" (durable) and "**AZIMAT**" (amulet), take place on words that contain characters which rarely use in Indonesian language like "W", and "Z". The distribution of the both characters in Indonesian words are included in the bottom part, so naturally it was hard to found a bi-gram that contains those characters in "Letter Shift" part of modified Baudot-Murray Code. Because of that reason, the encoding of the secret message contains "**444**" (Shift to Figure or Shift to Letter) causing a decrease of efficiency.

4.2 English Language Secret Message

Although the modified Baudot-Murray Code created using Indonesian language corpus, in this section the experiment was conducted to find out the performance of the improved NORMALS method if the secret message was written using English language (see Table 6).

Table 6 Experiments Using English Language

SECRET MESSAGE	PREVIOUS METHOD		PROPOSED METHOD		SAVING CAPACITY
	PREVIOUS METHOD	QUESTIONS NEEDED / REMAINING QUESTIONS	PROPOSED METHOD	QUESTIONS NEEDED / REMAINING QUESTIONS	
CAN	$01000011_2\ 01000001_2\ 01001110_2$	12 questions / 7 questions	$230_5\ 111_5$	6 questions / 13 questions	6 questions
FOR	$01000110_2\ 01001111_2\ 01010010_2$	12 questions / 7 questions	$444_5\ 303_5\ 424_5$	9 questions / 10 questions	3 questions
BANK	$01000010_2\ 01000001_2\ 01001110_2\ 01001011_2$	16 questions / 3 questions	$020_5\ 444_5\ 213_5$	9 questions / 10 questions	7 questions
ALARM	$01000001_2\ 01001100_2\ 01000001_2\ 01010010_2\ 01001101_2$	20 questions / -1 questions	$433_5\ 323_5\ 004_5$	9 questions / 10 questions	11 questions
PRAISE	$01010000_2\ 01010010_2\ 01000001_2\ 01001001_2\ 01010011_2\ 01000101_2$	24 questions / -5 questions	$444_5\ 300_5\ 444_5\ 400_5\ 422_5$	15 questions / 4 questions	9 questions
WITHOUT	$01010111_2\ 01001001_2\ 01010100_2\ 01001000_2\ 01001111_2\ 01010101_2\ 01010100_2$	28 questions / -9 questions	$444_5\ 001_5\ 431_5\ 404_5\ 421_5\ 432_5\ 431_5$	21 questions / -2 questions	7 questions

After analyzing the result of encoding the secret message in English language, the analysis shows that there are some bi-grams that commonly arise in both languages, English and Indonesian. For example, in the experiments, both "**CAN**" and "**ALARM**" showing a good result because "**CA**", "**AL**" and "**AR**" are common bigrams that found in Indonesian language while "**N** " and "**M** " are common last character of words in Indonesian language.

Whereas "**BANK**" another English word that already absorbed to Indonesian language show average result, because while "BA" is a common bigrams in Indonesian language, "NK" is not so common.

The worst results were obtained while encoding "**FRY**" and "**WORTHY**". Although there are bi-grams "**FR**" and "**Y** " and "**RT**" in the modified Baudot-Murray code, but they are located in the "**Figure Shift**" and need "**444**" (Shift to Figure) in order to use them. And for character combination "WO" and "HY", while the corpus analysis found that combination in the corpus but their appearance was too small to be included in the modified Baudot-Murray code so that the encoder has to use single character encoding.

Special note for "**WORTHY**" secret message, because the encoder has to alternate switching between "**Letter Shift**" and "**Figure Shift**" the actual result is worse than if the encoder uses single character encoding then it only needs 21 questions to hide the secret message instead of 24.

4.3 Secret Message with Non-Linguistic

To proceed to test the performance of the improved NORMALS method, the experiments continued using a set of random character that doesn't belong to linguistic rule (see Table 7).

Table 7 Experiments Using Non-Linguistic

SECRET MESSAGE	PREVIOUS METHOD		PROPOSED METHOD		SAVING CAPACITY
	PREVIOUS METHOD	QUESTIONS NEEDED / REMAINING QUESTIONS	PROPOSED METHOD	QUESTIONS NEEDED / REMAINING QUESTIONS	
ANDRE	$01000001_2\ 01001110_2\ 01000100_2\ 01010010_2\ 01000101_2$	20 questions / -1 questions	$000_5\ 444_5\ 400_5\ 424_5\ 401_5$	15 questions / 4 questions	5 questions
JOKOWI	$01001010_2\ 01001111_2\ 01001011_2\ 01001111_2\ 01010111_2\ 01001001_2$	24 questions / -5 questions	$444_5\ 411_5\ 421_5\ 444_5\ 231_5\ 444_5\ 001_5$	21 questions / -2 questions	3 questions
CYNTHIA	$01000011_2\ 01011001_2\ 01001110_2\ 01010100_2\ 01001000_2\ 01001001_2\ 01000001_2$	28 questions / -9 questions	$444_5\ 344_5\ 441_5\ 444_5\ 303_5\ 444_5\ 043_5\ 342_5$	24 questions / -5 questions	4 questions
GVPOTJ	$01000111_2\ 01010110_2\ 01010000_2\ 01001111_2\ 01010100_2\ 01001010_2$	24 questions / -5 questions	$444_5\ 403_5\ 433_5\ 2235\ 431_5\ 411_5$	18 questions / 1 questions	6 questions

Only a few bi-grams from analyzed corpus appear in "**ANDRE**", "**JOKOWI**" or **"CYNTHIA"**, worst of all the encoder could not found any bi-grams in the "**GVPOTJ**". In this scenario, the encoder often has to revert to use single character encoding instead of bi-grams.

4.4 Resume of Experiments

By considering that the modified Baudot-Murray table was created by using bi-grams analysis and for each code in the table need 3 questions to be embedded into the cover text, for the questionnaire with 19 questions that used in this experiments the minimum and maximum length of the secret message that could be embedded can be calculated:

1) Maximum embedding capacity for this experiments is 12 characters. This can be achieved if the secret message contains words which can be encoded using symbols contained in "**Letter Shift**" (bi-grams).

$$Maximum\ embedding\ capacity = \left\lfloor \frac{Number\ of\ questions}{3} \right\rfloor \times 2$$

$$Maximum\ embedding\ capacity = \left\lfloor \frac{19}{3} \right\rfloor \times 2 = \lfloor 6.33 \rfloor \times 2 = 12\ characters$$

While for NORMALS, the embedding capacity can be calculated as:

$$NORMALS\ embedding\ capacity = \left\lfloor \frac{Number\ of\ questions}{4} \right\rfloor$$

$$NORMALS\ embedding\ capacity = \left\lfloor \frac{19}{4} \right\rfloor = \lfloor 4.75 \rfloor = 4\ characters$$

2) For the worst case the improved NORMALS method can only embed 4 characters into the secret message.

$$Min.\ embedding\ capacity = \left\lfloor \frac{(Number\ of\ questions)}{15} \right\rfloor \times 4 + \left\lfloor \frac{(Number\ of\ questions)\ Mod\ 15}{6} \right\rfloor \times 1$$

$$Min.\ embedding\ capacity = \left\lfloor \frac{19}{15} \right\rfloor \times 4 + \left\lfloor \frac{19\ Mod\ 15}{6} \right\rfloor \times 1 = \lfloor 1.26 \rfloor \times 4 + \lfloor 0.67 \rfloor \times 1 = 4\ characters$$

Based on saving capacity, the improved NORMALS always show some improvement compared to NORMALS. Usually, a secret message written in Indonesian language give better saving capacity than English language, but it all depends on how the secret message was written. A secret message written using a pair of characters that often appear in the corpus will provide better saving capacity compared with a secret message written using the pair of characters that rarely appear.

Nevertheless, the secret message written in Indonesian and English language indicated better saving capacity while the secret message written in the non-linguistic shown worse saving capacity.

As a comparison, because the original NORMALS method using ASCII to encode the secret message, they always show the same result whether the secret messages are written in Indonesian language, in English language or even written with a random character.

5. CONCLUSIONS AND FUTURE WORK

Based on saving capacity, it has been proven that the improved NORMALS method has a better performance, thus has a better the embedding capacity compared to NORMALS.

The embedding capacity of improved NORMALS method depends on the compatibility between analyzed corpus used to generate modified Baudot-Murray code and the secret message. If the secret message consists of bi-grams contained in "Letter Shift" part of the modified Baudot-Murray table, then the improved NORMALS will show the best result.

One of the key performance measures of the steganography method is embedding capacity. Although improved NORMALS has proved to be better than NORMALS, in the worst scenario the increase is not significant. The embedding capacity of improved NORMALS can be further improved by searching a method that can reduce movement between "Shift to Letter" and "Shift to Figure" or using another corpus which has high compatibility with a secret message to be written.

6. ACKNOWLEDGMENTS

We thank our colleagues from Telkom University who provided insight and expertise that greatly assisted the research.

7. REFERENCES

[1] A. Desoky, "Nostega: A Novel Noiseless Steganography Paradigm," *Journal of Digital Forensic Practice*, vol. 2, no. 3, pp. 132-139, 2008.

[2] A. Desoky, "Graphstega: Graph Steganography Methodology," *Journal of Digital Forensic Practice*, vol. 2, no. 1, pp. 27-36, January 2008.

[3] A. Desoky, "Chestega: Chess Steganography Methodology," *Journal of Security and Communication Networks*, vol. 2, no. 6, pp. 555-566, March 2009.

[4] A. Desoky, "Edustega: An Education-Centric Steganography Methodology," *International Journal of Security and Networks*, vol. 6, no. 2/3, pp. 153-173, 2011.

[5] A. Desoky, "NORMALS: Normal Linguistic Steganography Methodology," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 1, no. 3, pp. 145-171, July 2010.

[6] D. Jurafsky dan J. H. Martin, Speech and Language Processing, 2nd Edition, Prentice Hall, 2008.

[7] E. Reiter dan R. Dale, Building Natural Language Generation Systems (Studies in Natural Language Processing), Cambridge University Press, 2006.