International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 19 (2015) pp 40209-40215

A Text-To-Speech using Rule-based and Data-driven Prosody Techniques with Concatenative Synthesis of the Philippines' *Bisaya* Dialect

R. A. Caroro

Professor College of Computer Studies Misamis University Ozamiz City, Misamis Occidental Philippines
claire130705@gmail.com

A. B. Garcia

Professor Department of Information Technology Occidental Mindoro State College San Jose, Occidental Mindoro Philippines
ailenc2004@yahoo.com

C. S. Namoco Jr.

Professor College of Industrial and Information Technology Mindanao University of Science and Technology Cagayan de Oro City, Philippines csnamocojr@yahoo.com

Abstract
The Philippines' *Bisaya* dialect is a Malayo-Polynesian-based language belonging to the Austronesian family of languages. It is the language spoken in the provinces of *Visayas* and the communicative and trade language in almost all of *Mindanao*, all in the Philippines. With the *Bisaya* language's popularity in the country and its widespread use, the development of the *Bisaya* text-to-speech (TTS) system help aspiring speakers to learn the language. This study aimed to enhance the existing *Bisaya* Text-To-Speech (TTS) system to produce a more natural and more intelligible speech. Known as EBiTTS or the Enhanced Bisaya Text-To-Speech system, it utilized rule-based and data-driven prosody techniques with concatenative synthesis. It analyzed simple *Bisaya* sentences and introduced prosodic features such as pitch shift and pitch duration. A listening assessment by 30 native *Bisaya*-speaking respondents compared the performance of EBiTTS in terms of naturalness and intelligibility with that of the existing *Bisaya* TTS. The results of the evaluation revealed that EBiTTS is more natural-sounding and more intelligible TTS system than the existing *Bisaya* TTS. These imply that EBiTTS produced an audible and not annoying speech and require only a little listening effort to understand the meaning of the synthesized speech.

Keywords: Text-to-speech, Data-driven TTS, Rule-based TTS, Concatenative Synthesis, *Bisaya*.

Introduction

Text-to-speech (TTS) analyzed, processed and transformed an input text into a speech signal [1, 2], to produce a synthetic speech. TTS applications used various techniques, like rule-based [1, 3], concatenative [1, 3, 4] and data-driven [5, 6]. Rule-based technique developed specialized rules to derive the synthetic speech [3]. It generates a robotic-sounding speech with no speech databases [1]. On the other hand, concatenative synthesis concatenates equivalent segments or acoustic units of the text inputs to synthesize the speech [1, 4]. It produces a more natural-sounding synthesis using unit selection synthesis, diphone synthesis and domain-specific synthesis [1, 3, 4]. However, adding a prosodic model enhances the generated speech. Prosodic model defined the duration, intonation and intensity patterns of speech [7] to produce a more intelligible and comprehensive speech synthesis. The presence of a prosodic marking may produce contrast or semantic context of the utterance, which includes pauses and prosodic phrasing, pitch, rate/relative duration and loudness [8]. An appropriate prosody embedded in speech synthesis enables the listeners to interpret the context of the discourse, whether it is syntactic or emphatic [9]. Furthermore, homograph or heteronym disambiguation analyzed the parts-of-speech of the input token [10, 11] or the sentence input to understand the dialog context [11].

Several studies on speech synthesis systems using various techniques modeled and developed TTS for various languages, including the *Filipino* language and some Philippine dialects. In the Philippines, the *Bisaya* dialect is a Malayo-Polynesian-based language belonging to the Austronesian family of languages [12]. It is the language spoken in the provinces of *Visayas* and the communicative and a trade language in almost all of *Mindanao* [13], all in the Philippines. Its grammar includes phonology-orthography correspondence or the sound-letter system, morphology or the word system, and syntax or the sentence system [12]. Its spelling depends on its sounds, thus a vowel plays a very significant role in the language. It contains three vowels a, i, and u, which later recognizes the vowels e and o because of foreign colonization. It has 15 consonants, b, k, d, g, h, l, m, n, ng, p, r, s, t, w, and y.

With the *Bisaya* language's popularity in the country and its widespread use, the development of the *Bisaya* text-to-speech (TTS) system help aspiring speakers to learn the language. [14] developed an existing *Bisaya* TTS system to address the need. The TTS system only performed limited capabilities like normalizing and tokenizing the *Bisaya* text inputs and used monotonic speech utterance with no prosodic features applied. The present study, the Enhanced *Bisaya* Text-To-Speech

(EBiTTS) synthesizer, enhanced the existing *Bisaya* TTS synthesis, which included prosodic features to basically emphasize the difference between a phrase, a declarative sentence, interrogative and exclamatory; as well as enhancing the utterance of the stressed words. Furthermore, its prosodic features disambiguated the heteronyms when used in the basic *Bisaya* sentence structure. The utterance is narrative, thus, the interrogative and exclamatory sentences are dependent only on the punctuation indicated at the end of the sentence.

Methods

A. The Architecture of EBiTTS Synthesizer

The EBiTTS synthesizer created a module that enhanced the existing syllabication and synthesis of the *Bisaya* text input. Its linguistic processing added a module to convert a single digit to its *Bisaya* word equivalent, determined the sentence type and the heteronym occurrences, and identified the basic sentence structure aside from the existing text normalization, tokenization and syllabication processes. It created a *Bisaya* language lexicon for details of prosodic characteristics and added prosody in the synthesis. Prosodic modeling modeled the process of inserting pauses and pitch shift. It started with the selection of acoustic units, and then followed by shifting the pitch and then altering speech duration. Lastly, the generation of the speech followed, using concatenative synthesis. Figure 1 presents the structure of EBiTTS synthesizer, while Table 1 presents the conceptual comparison of the existing *Bisaya* TTS system [14] and the enhanced system.

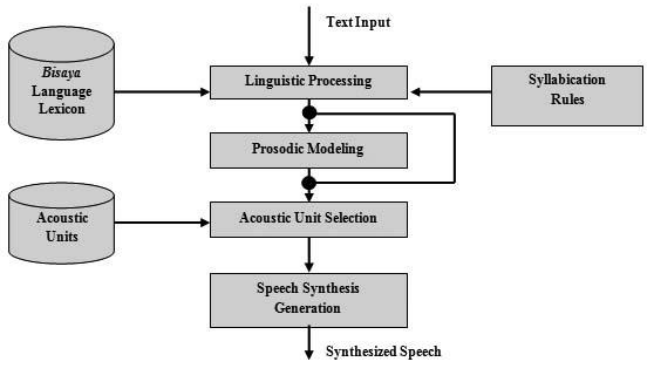


Fig. 1. The General Architecture of EBiTTS synthesizer

B. Linguistic Processing

The linguistic processing module covered a number of sub-processes, including text normalization and tokenization, syllabication, basic sentence structure and heteronym ambiguity analysis, as presented in Figure 2. It first normalizes the text by removing invalid characters like series of digits including a date entry, and some symbols. However, a single digit token, punctuations like period, question mark, exclamation point, comma, and hyphen, as well as the whitespace character are valid.

Linguistic processing then tokenized the normalized text inputs. The text tokenization step used the whitespace as a delimiter to mark the separation of tokens. In order to retrieve the digit's acoustic equivalent; the module converts the single digit into its equivalent *Bisaya* word. Syllabication determined syllable boundaries to retrieve the equivalent phones [15, 16]. The syllabication process of the study used the syllabication rules identified by [14], presented in Table 2. The syllabication process used vowels to extract the syllables and recognized and syllabicated hyphenated words.

Table 1. Conceptual Comparison of the Existing *Bisaya* TTS [14] system and EBiTTS

Featured Processes	Existing *Bisaya* TTS System	EBiTTS
Text normalization	✓	✓
Text tokenization	✓	✓
Single-digit to word conversion		✓
Syllabication without hyphen	✓	✓
Syllabication with hyphen		✓
Sentence type analysis		✓
Heteronym ambiguity analysis		✓
Basic sentence structure analysis		✓
Pitch shifting for stresses		✓
Lengthening the duration (time in seconds) for pause		✓
Acoustic unit selection	✓	✓
Speech synthesis	✓	✓

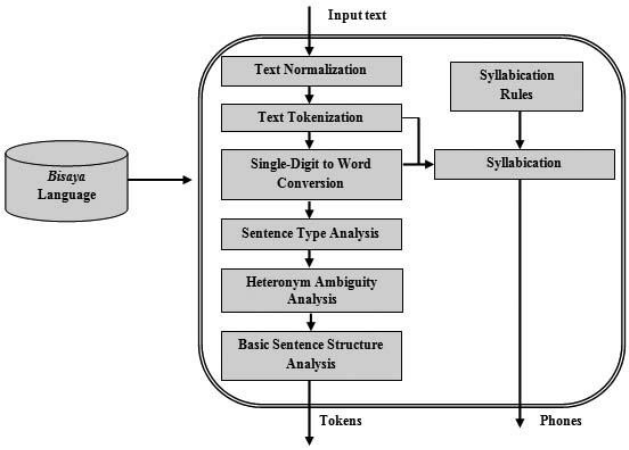


Fig. 2. Block Diagram Model of the Linguistic Processing Procedure

The study determined the type of sentence of the text input as a declarative statement, interrogative or exclamatory using the end-of-the-sentence punctuation as the delimiter. The period or its absence at the end of the text input determined a declarative sentence; the question mark for the interrogative sentence; and the exclamation point for the exclamatory sentence. A comma in the sentence determined the location of a pause during synthesis. The module shifts the pitch of these types of sentences however, the synthesis of these types of sentences only sound narrative.

Table 2. Syllabication Rules [14]

	Rule	Example
1	Vowel alone - one syllable	*"i" + "kaw"*
2	Vowel + *"ng"* - one syllable	*"a" + "ng"*
3	Vowel + Consonant - one syllable	*"a" + "b" + "ri"*
4	Consonant + Vowel - one syllable	*"b" + "a" + "ta"*
5	Consonant + Vowel + Consonant - one syllable	*"h" + "a" + "g" + "dan"*
6	Consonant + Vowel + Vowel - two syllables (Consonant-Vowel and Vowel alone)	*"m" + "a" + "a" + "yo"*
7	Consonant + Vowel + *"ng"* - one syllable	*"b" + "i" + "ng" + "ka"*
8	Consonant + Consonant + Vowel – one syllable	*"k" + "l" + "a" + "ro"*
9	*"ng"* + Vowel – one syllable	*"ng" + "a" + "lan"*
10	*"ng"* + Vowel + Consonant – one syllable	*"ng" + "a" + "n" + "hi"*
11	Consonant + Consonant + Vowel + *"ng"* – one syllable	*"p" + "l" + "a" + "ng" + "ka"*

For text input having heteronyms, heteronym ambiguity analysis determined its occurrence. A heteronym is a word having the same spelling with various meanings and pronunciation [17]. The *Bisaya* language lexicon determined whether the token or word is a heteronym by checking word matches in the lexicon. A priority marker, numbered as 1, 2, and 3, retrieved the word in the lexicon. All words which are not heteronym have priority marker equal to 1. However, a priority marker from 1 to n determined the total occurrences of that word in the lexicon. Upon retrieving the prosodic characteristics of the word input, heteronym ambiguity analysis used the priority marker equal to 1 to retrieve a single word input which is a heteronym. Priority markers equal to 1 and 2 retrieved the word which appears twice in a three-word sentence input where the second word is a noun phrase marker. Likewise, a priority marker equal to 1 retrieved a heteronym in a two-word input and in longer sentence inputs. The linguistic processing module then retrieved the specific word match and its equivalent prosodic characteristics.

Basic sentence analysis used the sentence structure or sentencing pattern of [12] considering also the noun phrase markers identified by [18] in order to determine the part-of-speech for a certain word. Table 3 presents the *Bisaya* simple sentence patterns using the given grammar rules, using the notations n (noun), pm (phrase marker), v (verb), adj (adjective) and adv (adverb). *Bisaya* phrase markers determined the subject of the sentence and its relation to the predicate. The basic sentence structure analysis traced the tokens from the text inputs to determine a heteronym occurrence.

Table 3. Basic *Bisaya* Sentence Grammar Rules

Sentence Pattern	Description	Example
Noun + Noun	A noun is followed by a noun phrase marker then by another noun	*"Puthaw ang haligi."* (n) (pm) (n)
Adjective + Noun	An adjective is followed by a noun phrase marker then by a noun	*"Humut ang bulak."* (adj) (pm) (n)
Adverb + Noun	An adverb is followed by a noun phrase marker then by a noun	*"Kadugay sa bayad."* (adv) (pm) (n)
Verb + Noun	An verb is followed by a noun phrase marker then by a subject	*"Nagbasa ang bata."* (v) (pm) (n)

C. Prosodic Modeling

The prosodic modeling phase focused on the design of the prosodic model to present the prosodic features in the speech synthesis. Prosodic features scaled the frequency and time domains of the audio signals on stresses, and duration lengthening on pauses. Figure 3 presents the prosodic model of EBiTTS.

The modeling process added the prosodic features to the equivalent acoustic units through a series of defined algorithms. First is determining whether the token is in the lexicon followed by determining the stress locations on the syllables, then by determining the punctuation at the end of the sentence, and by adding the prosodic features to the equivalent audio file. In addition, prosodic modeling searched the occurrences of comma in the input word and added 0.5 seconds duration to the synthesis before the retrieval and the playing of the next audio signal.

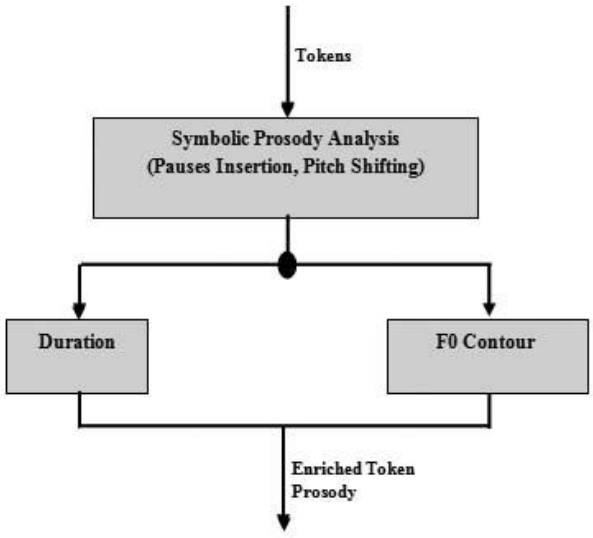


Fig. 3. Prosodic Model of EBiTTS

The prosodic model shifted the pitch of the stressed and punctuated words. Symbolic prosodic analysis lengthened the duration of the audio signal to provide pauses in the sentence based on the position of the comma.

The symbolic prosody process also determined the final token or syllable having the pitch shift on interrogative and exclamatory sentences. The pitch shifting algorithm raised the pitch of the final token or the last syllable of an interrogative sentence to indicate questioning. On the other hand, the algorithm raised the pitch of the last word of an exclamatory sentence and then dropped to indicate intense utterance. The study used a plain narrative utterance and ruled out other speaking styles or emotional conditions, such as amusement, anger, and other emotions having an effect in prosody. Prosodic features like pitch shifting and speech duration were dependent on the conditions identified by series of rules defined in the algorithm.

D. Acoustic Unit Selection

The process of unit selection used the analyzed text or phones in the previous modules and selected the acoustic units for speech synthesis. The selection retrieved the tokens having equivalent acoustic units, manipulated the prosodic characteristics to shift the pitch and to add a pause. A blank audio file replaced the instance of a syllable that does not have any equivalent acoustic unit [19, 20, 21]. This approach enabled the synthesis to continue despite having no matched acoustic unit on some syllables.

E. Speech Synthesis Generation

The speech synthesis generation module concatenated the selected acoustic units based on their order of occurrence in the text input and added the identified prosodic features. The study used the phase vocoder algorithm [22, 23, 24, 25] in scaling the frequency and time domain of the audio signals. It also lengthened the duration or time of the tokens preceded by commas with 0.5 seconds of the utterance to indicate pauses.

F. Bisaya Lexicon Development

The study developed a *Bisaya* language lexicon to serve as a resource for the text, phonetic and prosodic analysis, containing commonly used words by the *Bisaya*-speaking people of *Mindanao*. The study used the *Bisaya* words listed in [26]. Figure 4 presents the design of the *Bisaya* lexicon.

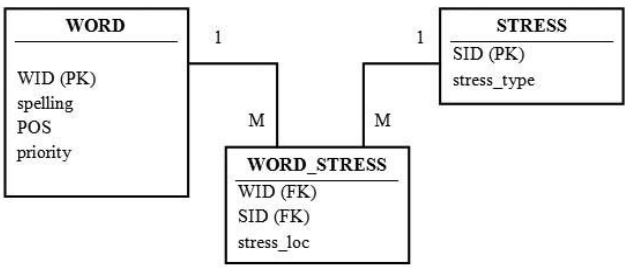


Fig. 4. The *Bisaya* Lexicon of EBiTTS

G. Speech Database (Acoustic Unit) Definition

Recording of acoustic units to make up the speech database ensures the coverage of the different frequencies of the syllables [5]. The study recorded the acoustic units per syllable and saved as a WAV file. A native *Bisaya* speaker read the syllables for recording. The recording took place in a semi-soundproof booth of a local FM radio station to minimize the noise interference that might occur during the recording. The study then split the recorded audio accordingly into acoustic units representing each syllable. Labeling then took place to mark the acoustic units using the syllable name as the file name.

H. Assessing Speech Synthesis' Naturalness and Inteligibility

The study assessed the performance of EBiTTS synthesizer according to naturalness and intelligibility [8, 27] in which having higher intelligibility result contributes to greater comprehensibility performance [28]. It conducted testing using single words representing the three types of stresses (acute, grave and circumflex), single words with hyphen, single-digit input, two-word inputs with punctuations, and longer sentences/phrases. The testing process used randomly-selected words to represent similar words from the lexicon having the same prosodic characteristics, and used them as single word input, phrases or sentence inputs. Each test used three samples for each input type. The test compared the performance of EBiTTS with the performance of the existing *Bisaya* TTS of [14].

Thirty (30) randomly-selected native *Bisaya*-speaking respondents assessed the naturalness and intelligibility of the speech synthesis generated by EBiTTS synthesizer. The same respondents assessed the naturalness and intelligibility of the existing Bisaya TTS system [14].

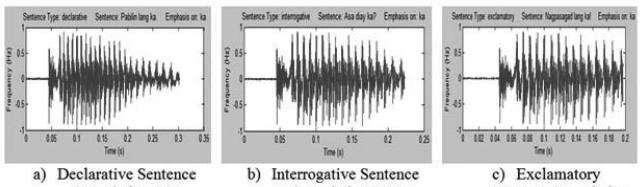
The respondents gathered to listen and rate EBiTTS performance in a listening assessment, followed by the testing of the existing *Bisaya* TTS [14]. The testing for both TTS synthesizers used the same text inputs and used the subjective Mean Opinion Score (MOS) listening only rubric [8, 27, 29] to analyze the naturalness and intelligibility of the two TTS synthesizers.

Results

A. The EBiTTS Synthesizer

The EBiTTS synthesizer normalized and tokenized the text inputs. It converted a single digit token into its equivalent *Bisaya* word and then syllabicated. It added a pitch shift to the last syllable of the input words having end-of-the-sentence punctuations, like a question mark and exclamation point. With the sampling frequency of 44100 Hertz, the duration or the time in seconds at which the generated audio signal is being sampled varies depending on the audio modification made for inputs having end-of-sentence punctuations. This made the pitch shifting process to synthesize the input sentence meaningfully to convey the sense of the sentence content as a whole and not literally. Figure 5 presents the audio signals generated by the synthesizer for the three different sentence types, while Figure 6 presents the sample audio signals of heteronyms. The synthesis showed that if

both occurrences appear in the same basic sentence, it selected the appropriate prosodic characteristics based on the rules for basic sentence analysis.



a) Declarative Sentence (sampled at 0.30 seconds)
b) Interrogative Sentence (sampled at 0.22 seconds)
c) Exclamatory Sentence (sampled at 0.197 seconds)

Fig. 5. Sample Audio Signals Generated for the Three Types of Sentences with Emphasis on *"ka"*: (a)*"Pabilin lang ka."*; (b)*"Asa diay ka?"*; (c)*"Nagpasagad lang ka!"*

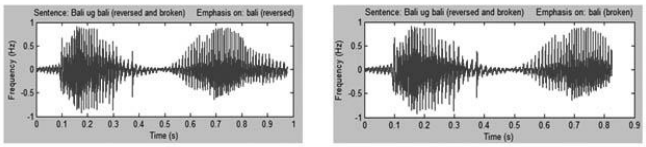
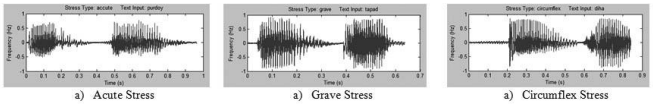


Fig. 6. Sample Audio Signals Generated for a Heteronym: "*balí ug balî*" (left and right sample audio signals, respectively)

EBiTTS determined the stresses on text inputs and applied the prosodic characteristics to the input. It searched and matched a text input with the recorded words in the lexicon to retrieve the prosodic characteristics. After which, it concatenated the equivalent audio files for the syllables in the input text, forming a single speech and played using the common frequency sample of 44100 Hertz. Each audio equivalent for the syllable synthesized at lower amplitude before playing a new audio file for the next syllable. EBiTTS produced synthesis of stressed syllables which sounded differently from the unstressed syllables. Figure 7 presents sample audio signals generated by the synthesizer according to the type of stress the input has.



a) Acute Stress
a) Grave Stress
a) Circumflex Stress

Fig. 7. Sample Audio Signals Generated for the Three Types of Stresses: (a)*"púrdoy"*; (b)*"tapàd"*; (c)*"dihâ"*

B. The Evaluation Process

A listening assessment of 24 random sets of text inputs evaluated the naturalness and intelligibility of the existing *Bisaya* TTS [14] and EBiTTS. The result showed that the existing *Bisaya* TTS [14] is fair (MOS = 2.84, SD = 1.27), which the speech is slightly annoying and some errors in the utterance were noticeable. Likewise, the naturalness of EBiTTS is good (MOS = 3.94, SD = 0.19), which the speech is audible and not annoying even if some speech errors were slightly noticeable. Table 4 presents the results of the evaluation on naturalness of the two TTS systems.

Table 4. Mean Opinion Score (MOS) and Standard Deviation (SD) for Naturalness at Different Input Types Using the *Bisaya* TTS[14] and EBiTTS

TTS Systems \ Input Types		Single word – acute stress	Single word – grave stress	Single word – circumflex stress	Single word with hyphen	Single digit	Basic sentence with heteronym	Two word input with punctuations	Longer sentences with punctuations	Overall MOS
Bisaya TTS (Garcia, et al., 2015)	MOS	4.09	4.06	3.60	1.00	1.00	3.68	2.98	2.32	**2.84**
	(SD)	(0.13)	(0.35)	(0.23)	(0.00)	(0.00)	(0.65)	(0.86)	(0.31)	**(1.27)**
EBiTTS	MOS	4.12	3.80	3.80	4.26	3.88	3.78	3.78	4.13	**3.94**
	(SD)	(0.05)	(0.45)	(0.20)	(0.05)	(0.10)	(0.08)	(0.24)	(0.10)	**(0.19)**

The existing *Bisaya* TTS system [14] commonly produced a natural sounding speech on single word inputs of varying stresses because of the per-syllable recording and playing, which make each syllable clear to the listeners. However, the system is not capable of synthesizing inputs having hyphen and other punctuations as well as single digits. When the synthesizer encountered an input with punctuations, the synthesis skipped those words making the synthesis incomplete. On the other hand, the system can synthesize longer sentences or phrases. However, the system is unable to present a considerable gap in the utterance thus producing speech synthesis with overlapping speeches. The result showed that the existing *Bisaya* TTS needs further normalization to cater these system needs. On the other hand, EBiTTS can synthesize all word inputs with or without hyphen or punctuations. This enabled the system to recognize all inputs and to produce the equivalent speech synthesis. Furthermore, the synthesis produced a speech with considerable gaps in between word utterance, as well as words having comma. This made EBiTTS to produce a complete speech with considerable gaps and no missing words. The result also showed that with the comprehensive normalization process made realized the naturalness of the speech synthesis. The use of syllables as basis for concatenation and the prosodic matching used in the data-driven concatenative synthesis supported these capabilities. Furthermore, the *Bisaya* lexicon used in EBiTTS helped in achieving a natural-sounding synthesis.

Further, the listening assessment evaluated the existing *Bisaya* TTS [14] and EBiTTS' intelligibility and showed that the existing *Bisaya* TTS [14] is fair (MOS = 2.78, SD = 1.27), which the speech requires moderate listening effort to understand the meaning of the synthesis. Likewise, the naturalness of EBiTTS is good (MOS = 3.93, SD = 0.22), which the speech requires a little listening effort to understand the meaning of the synthesis. Table 5 presents the results of the evaluation on intelligibility of the two TTS systems.

Table 5. Mean Opinion Score (MOS) and Standard Deviation (SD) for Intelligibility at Different Input Types Using the *Bisaya* TTS[14] and EBiTTS

TTS Systems	Input Types	Single word – acute stress	Single word – grave stress	Single word – circumflex stress	Single word with hyphen	Single digit	Basic sentence with heteronym	Two word input with punctuations	Longer sentences with punctuations	Overall MOS
Bisaya TTS (Garcia, et al., 2015)	MOS	4.10	4.00	3.54	1.00	1.00	3.59	2.84	2.08	**2.78**
	(SD)	(0.02)	(0.25)	(0.35)	(0.00)	(0.00)	(0.57)	(0.98)	(0.39)	**(1.27)**
EBiTTS	MOS	3.92	3.73	3.86	4.27	3.86	3.72	3.81	4.26	**3.93**
	(SD)	(0.05)	(0.44)	(0.16)	(0.15)	(0.18)	(0.12)	(0.27)	(0.11)	**(0.22)**

The single word input with acute stress in the existing *Bisaya* TTS [14] has the highest rate and interpreted as good, whereas the single word with a hyphen and single digit inputs has the lowest rate and interpreted as bad. Meanwhile, the synthesis of single word with hyphen had the highest rated input in EBiTTS and interpreted as excellent, whereas the inputs on basic sentence with heteronym have the lowest rating but still interpreted as good. The existing *Bisaya* TTS system [14] produced a synthesis having intelligible context on single word inputs only. However, if the synthesis contained longer phrases or sentences, the synthesis became poorly intelligible because of overlapping speeches and no consistent considerable gaps in between word utterance. In addition, missing synthesis is evident in the existing *Bisaya* TTS [14] when the inputs have punctuations or single digit. The result showed that the existing *Bisaya* TTS is unable to recognize and normalize inputs with punctuations. Furthermore, it is unable to provide a consistent gap in between word utterance to produce speech synthesis with considerable gaps. On the other hand, EBiTTS can synthesize all word inputs with or without hyphen or punctuations. This enabled the system to recognize all inputs and to produce the equivalent speech synthesis. Furthermore, the synthesis produced a speech with considerable gaps in between word utterance, as well as words having comma which produced a considerable pause before the next word utterance. This made EBiTTS to produce an intelligible speech utterance with considerable gaps and no missing word synthesis. The result also showed that the comprehensive normalization process of EBiTTS made the intelligibility of the speech synthesis realized. Furthermore, EBiTTS produced a complete and intelligible synthesis, which conveyed the real meaning of the sentence.

Conclusion

The enhanced *Bisaya* text-to-speech technology (EBiTTS) used rules for extracting syllables and for basic sentence analysis; *Bisaya* lexicon for prosodic characteristics of the *Bisaya* words and concatenative approach for speech synthesis. EBiTTS considered the vowels as significant indicators for the number of syllables composing a given word. In addition, EBiTTTS used the concept of the *Bisaya* language's basic sentence which always ends with a noun, thus EBiTTS identified first whether a heteronym, as the third token in a sentence, is a noun, before determining other parts of speech.

Evaluation on speech naturalness of the existing *Bisaya* TTS revealed that it is fair, which speech quality is slightly annoying and some errors in the utterance were noticeable. Likewise, EBiTTS' naturalness result revealed that it is good, which the speech is audible and not annoying even if some speech errors were slightly noticeable. On the other hand, the evaluation on intelligibility of the existing *Bisaya* TTS showed that the system is fair, which the speech requires moderate listening effort to understand the meaning of the synthesis. Likewise, the evaluation on intelligibility of EBiTTS showed that the synthesizer is good, which the speech requires a little listening effort to understand the meaning of the synthesis.

The results further revealed that all the respondents choose EBiTTS as the good-sounding synthesizer in terms of naturalness and intelligibility than the existing *Bisaya* TTS.

Future researches further enhance EBiTTS to include widening the scope of sentence analysis to cover complex sentences of the *Bisaya* language; considering other prosody features such as speaking styles or speech emotions in the TTS system; and considering the possibility of recording a *Bisaya* speech (rather than per syllable recording) that is long enough to represent almost all syllable utterances to produce a more articulated speech.

References

[1] A. Chauhan, V. Chauhan, G. Singh, C. Choudhary, and P. Arya, "Design and development of a text-to-speech synthesizer system," *International Journal of Electronics and Communication Technology*, vol. 2, no. 3, pp. 42, 44, 2011.

[2] N. P. Narendra, K. S. Rao, K. Ghosh, R. R. Vempada, and S. Maity, "Development of syllable-based text to speech synthesis system in Bengali," *International Journal of Speech Technology*, vol. 14, no. 3, pp. 167-181, 2011.

[3] O. O. Khalifa, Z. H. Ahmad, A. H. Hashim, and T. S. Gunawan, "SMaTalk: Standard Malay text to speech talk system," *Signal Processing: An International Journal*, vol. 2, no. 5, p. 1, 2008.

[4] M. S. Shirbahadurkar, and D. S. Bormane, "Speech synthesizer using concatenative synthesis strategy for Marathi language (spoken in Maharashtra, India)," *International Journal on Recent Trends in Engineering*, vol. 2, pp. 80-82, 2009.

[5] S. P. Kishore, R. Kumar, and R. Sangal, "A data driven synthesis approach for indian languages using syllable as basic unit," In Proceedings of *Intl. Conf. on NLP (ICON)*, pp. 311-316, 2002.

[6] F. Tesser, P. Cosi, C. Drioli, and G. Tisato, "Prosodic data driven modelling of a narrative style in Festival TTS," In *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[7] V. R. Reddy, and K. S. Rao, "Intonation modeling using linguistic, production and prosodic constraints for syllable based TTS systems," *Procedia Engineering*, vol. 38, pp. 2772-2783, 2012.

[8] E. Morais, and F. Violaro, "Data-driven text-to-speech synthesis," *XXII Simpósio Brasileiro de Telecomunicações*, Campinas, Brazil, 2005.

[9] J. A. Erekson, "Prosody and interpretation," *Reading Horizons*, vol. 50, no. 2, p. 3, 2010.

[10] V. Tesprasit, P. Charoenpornsawat, and V. Sornlertlamvanich, "A context-sensitive homograph disambiguation in Thai text-to-speech synthesis," In Proceedings of the *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003—short papers*, vol. 2, pp. 103-105, 2003. Association of Computational Linguistics.

[11] O. Gago, S. M. Hancock, and M. E. Smith, "Disambiguating text that is to be converted to speech using configurable lexeme based rules," U.S. Patent No. 8,538,743, Sept. 17, 2013.

[12] A. Pesirla, "A pedagogic grammar for cebuano-visayan," In *1st MLE Conference, "Reclaiming the Right to Learn in One's Own Language*, 2010.

[13] D. A. P. Endriga, "The dialectology of Cebuano: Bohol, Cebu and Davao," In *1st MLE Conference, "Reclaiming the Right to Learn in One's Own Language*, pp. 18 – 20, 2010.

[14] A. B. Garcia, J. N. N. Rojo, and C. S. Namoco Jr., "A bisaya text-to-speech (TTS) system utilizing rule-based algorithm and concatenative speech synthesis," *European Academic Research*, vol. 2, no. 10, pp. 12997-13012, 2015.

[15] L. Narupiyakul, A. Khumya, B. Sirinaovakul, and N. Cercone, "A stochastic knowledge-based Thai text-to-speech system," *Mathematical and Computer Modeling*, vol. 42, no. 1-2, pp. 1-16, 2005.

[16] S. Mohanty, "Syllable based Indian language text to speech system," *International Journal of Advances in Engineering & Technology*, vol. 1, no. 2, pp. 138-143, 2011.

[17] J. Li, D. Xu, L. Yi, X. Lou, J. Luan, X. Wang,... and J. Hao, "The Toshiba Mandarin TTS system for the Blizzard Challenge 2008," *Challenge*, vol. 3, 2008.

[18] A. T. Parnes, "Ang marks the what?: An analysis of noun phrase markers in Cebuano," Yale University, 2011. [Online]. Available: http://ling.yale.edu/sites/default/files /alumni%20senior%20essays/Ava%20Tattleman%20 Parnes.pdf. [Accessed: 15-Jan-2015].

[19] M. C. Beutnagel, M. Mohri, and M. D. Riley, "Methods and apparatus for rapid acoustic unit selection from a large speech corpus," U.S. Patent No. 7,082,396 B1, July 25, 2006.

[20] M. C. Beutnagel, M. Mohri, and M. D. Riley, "Methods and apparatus for rapid acoustic unit selection from a large speech corpus," U.S. Patent No. 7,369,994 B1, May 6, 2008.

[21] M. C. Beutnagel, M. Mohri, and M. D. Riley, "Speech synthesis from acoustic units with default values of concatenation cost," U.S. Patent No. 8788268 B2, July 22, 2014.

[22] J. L. Flanagan, and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493-1509, 1966.

[23] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 3, pp. 243-248, 1976.

[24] D. Ellis, "A phase vocoder in Matlab," *Ee.columbia.edu*, 2010. [Online]. Available: http://www.ee.columbia.edu/ln/rosa/matlab/pvoc/. [Accessed: 05-Feb-2015].

[25] K. S. Rao, "Prosody knowledge for speech systems: a review," In *Predicting Prosody from Text for Text-to-Speech Synthesis*, pp. 7-25, 2012. Springer, New York.

[26] Cebu.sandayong.com, "Bisaya Translator and Cebuano Dictionary," 2012. [Online]. Available: http://cebu.sandayong.com/dictionary.aspx. [Accessed: 28-Jan-2015].

[27] D. J. Ravi, and S. Patilkulkarni, "Evaluation of Kannada text-to-speech [KTTS] system," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 1, 2012.

[28] Y. Y. Chang, "Evaluation of tts systems in intelligibility and comprehension tasks," In Proceedings of the *23rd Conference on Computational Linguistics and Speech Processing*, pp. 64-78, 2011. Association of Computational Linguistics.

[29] I. Rec, "P. 800.1, mean opinion score (MOS) terminology," *International Telecommunication Union, Geneva*, 2006.