Emotional Techy Basyang: An Automated Filipino Narrative Storyteller

John Christopher P. Gonzaga, Jemimah A. Seguerra, Jhonnel A. Turingan, Mel Patrick A. Ulit, and Ria A. Sagum

***Abstract*—This study is specifically concerned in developing a storyteller application which uses sentiment analysis and includes Text-to-Speech (TTS) that converts the input text story into its audio output. Concatenative synthesis was the algorithm used in the TTS process wherein every speech audio that represents each syllables will be concatenated to each other with some pauses for speech turns and delimiters. Also, the researchers used N-Gram in order to syllabicate every word in the story input. This study is aimed to determine its acceptability to the users and its accuracy in terms of its audio modification. The accuracy of the said application is measured by precision and its error rate.**

***Index Terms*—Sentiment analysis, text-to-speech, text categorization, tagalog stemmer**

I. Introduction

A storyteller is the one that narrates stories to its audience. An effective storyteller narrates with emotions that will help the audience to have more insights about the characters in the story. Stories nowadays can be narrated through technologies such as computers and creating storyteller application.

There are already systems in our technology in these days that can narrate stories but the problem is doing the narration to be natural or assigning voices in every character is quiet hard. In other words, the existing systems about storytelling can't fully generate human-like synthesized voice of characters. Synthesized voice narration can be accomplished by the process of text-to-speech (TTS) synthesis. That is, transforming written text into speech that people can listen to and understand. However, the voice produced by synthesizers often sound monotonous and artificial, thus causing it to be rejected by users. The difficulty lies in trying to express emotion in the synthesized speech to make it sound more natural and human-like.

The study aimed to develop a storyteller application that assigns appropriate voices with appropriate emotions in every situation of the character in the story. Applying right pause durations between sentences to distinguish the changes of scene and speech turns in the story are also included.

Manuscript received February 7, 2014; revised April 8, 2014. This work was supported in part by the Polytechnic University of the Philippines. Emotional Techy Basyang: An Automated Filipino Storyteller.

The authors are with the Department of Computer Science, Polytechnic University of the Philippines (e-mail: riasagum31@yahoo.com, gonzaga_jc1993@yahoo.com, jemimah.seguerra@gmail.com, jhonnelturingan@gmail.com, mlptrckulit@gmail.com).

II. Related Works

The researchers focused in solving the problem of how can storytellers implement a text-to-speech with a humanlike sound [1].

The main goal of the researchers is not just simply producing humanlike sound, but it also includes the adding of emotions to the speech output. To achieve their goal, the researchers used the following processes.

A. Used Festival Speech Synthesis System as Their Text-to-Speech Engine.

B. In emotion Recognition, the Researchers used Two Methods.

1) Employment of an actor

A professional actor is able to simulate emotions in his/her speech to such a degree that it is mostly indistinguishable from 'normal' spontaneous emotional speech.

2) Using a corpus of spontaneous emotional speech

The corpus was consists of speech fragments from a diverse range of people.

But even the researchers used these tools and algorithm, their result was not good enough just like the results of the majority of the researchers in that specific field. The TTS in the system needs to enhance its quality of voices and changing of intonation.

A research [2] also conducted a study about this application of Natural language (NLP). In their research in emotional storytelling, they realized that adding more emotions in it will make it more interesting. The researchers also introduced one way of recognizing emotion as shown in Fig. 1 which is through a modeled emotional path for each emotion category through the story, and an internal emotion tracking system, trying to predict the current emotional state of the story.

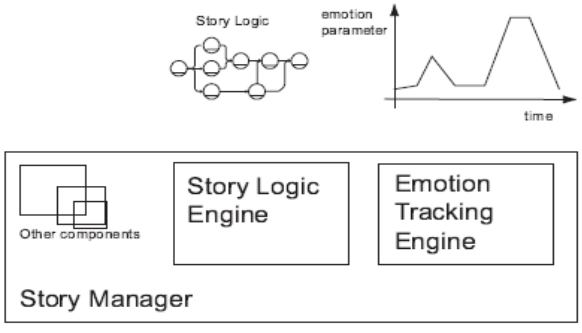

Fig. 1. Story manager with emotion tracking engine and emotional path. graph.

DOI: 10.7763/IJFCC.2014.V3.310

With the proposed changes, the IS (Interactive Storytelling) system has to have some changes to the algorithm the story manager follows. The new algorithm will then be as follows:

1) Using current IS system techniques for selection of logically possible story segments (from story logic)
2) From these choose based on (a best fit algorithm)
   - Time (from system)
   - Current user emotion (from Emotional Tracking Engine)
   - Emotional story logic
   - Values of the available story segments
3) Update the Emotional Tracking Engine with the values from the chosen story segment

The researchers feel that their proposed extensions to create Emotional Storytelling can be a great help to the area of interactive storytelling used for example in Virtual Reality. More importantly, the researchers believe that this emotional extension has various benefits. The largest benefit the researchers believe it provides is that of improved control by the author over creating an emotional story arc for the on-line generated, interactive story.

Another study was conducted by Susanne Hendrickx [3] also in translation of the emotions determined by the emotion model into natural language or speech. The researchers made an evaluation on the emotion model of the existing virtual storyteller using corpus study. As an approach to the problem, the researchers used corpus study to verify whether the emotions that are included in the emotion model match the emotions in the fairy tales. Based on the result, the researchers proposed some adjustments to the emotion model of the storyteller and identified the lexical items such as gradable adjectives to express emotions. After the lexical items of the emotions model, the researchers know how to refer to the two ends of an emotion scale. Next is to have a theoretical research on contrast relation that is capable of relating the expressed emotions of the story.

In the study entitled "Fairy Tales for Grown Ups" [4], the researchers attempted to provide a synthesizer with more understanding of its input in a shallow approach. The researchers aimed to have a more natural like output of children stories such as identifying the speech quotes, its speaker, assigning emotions to each clause and implement voice acting. For further researches, the researchers suggested to solve the problems such as the model doesn't perform analysis of other co-referential expressions and familiarity. The researchers also recommended having a database with predetermined sounds. The storyteller is also unclear how to deal with breaks and emphases. Finally, the mood classifier does not take into account the context of clause which would add more precision.

In the study entitled "Prosodic Analysis of a Corpus of Tales" [5], providing storytelling capacities to a humanoid robot is the main problem. This problem includes the improvement of text-to-speech synthesis expressivity according to a semi-automatic analysis of a given tale. To solve this problem, automatic tagging and prosodic stylization were applied to the corpus. The extracted parameters are described and analyzed according to relevant elements of the tales' structure. The results show that changes of pitch, use of devoicing are relevant to impersonate characters and to modify expressivity according to the different structural parts of each tale. The proposed approach extracts a set of global prosodic parameters from a large corpus of read tales, and linked them to a set of qualitative descriptors labeling the tales' linguistic and narrative analysis. These prosodic descriptions are then used to drive the synthesizer in its units-selection task. After presenting the textual and audio tales corpora, their linguistic and prosodic analysis is described. Then, a set of possibly relevant prosodic variations for enhancing the expressiveness of synthetic speech in accordance with a tale's narrative structure are listed and discussed.

In the research entitled "Emotion Recognition from Text Using Knowledge-based ANN" [6], the problem that the researchers want to answer is if they would be able to infer emotional state through implicit knowledge with the third-party information. The researchers used a system called KBANN or Knowledge Based Artificial Neural Network and Keyword based approach to get emotion from a sentence. Keyword based approach generated a 90% accuracy in getting emotion while KBANN only generated 45-65% accuracy in getting the emotions of sentences. The emotions that are gathered or being tested are Anger, Fear, Hope, Sadness, Happiness, Love, Thank, and Neutral.

The researchers had agreed on the problem of the development and assessment of a simple algorithm to interpolate the intended vocal effort in existing databases in order to create new databases with intermediate levels of vocal effort [7]. Interpolation is performed through linear prediction (LP) analysis/re-synthesis and uses the line spectral frequencies (LSFs) that are commonly employed in speech coding and voice conversion applications. For each pair of voice qualities, the interpolation factor is set to 0.5 in corresponding to the averaging of spectral envelopes and two interpolation results are obtained by using each of the residual signals without further modification or mixing. In the soft-modal interpolation examples, soft-to-modal refers to the version where residual of the soft database is utilized and modal-to-soft refers to the version where residual of the modal database is utilized for re-synthesis. The proponents have shown that both recorded and interpolated databases are perceived, quite independently of the language background of listeners and the phonetic string produced. Further research is necessary regarding the respective roles of the residual and the mixing ratio for interpolated voice quality.

III. Methodologies

A. Sentiment Analysis

Sentiment analysis is used to determine the present emotion in the detected speech quoted phrase. Sentiment analysis should undergo text categorization before it starts its analysis part.

B. Text Categorization

Text categorization is the task of determining the part of speech of a word. In this algorithm, some frequently used words will be used as guide lines in categorizing the words.

The following are examples of frequently used words:

1) Determiners

It is used as a signal word to point the word being referenced.

Example: Mga, Si, Sina, Ang

2) Linking verbs

It usually follows a noun and is followed by an adjective or verb.

Example: Ay

3) Other words such as pronouns, prepositions and numbers

4) Position parameter

The researchers included parameters that will help in determining the categorization of the detected word as a noun.

The position of the word in the sentence is necessary in determining its category. The algorithm includes the checking of the surrounding words and features of the words in a two-word-window. This means that the two words before and after the target word were used as added information to determine its category.

C. Tagalog Stemmer

The researchers used Tagalog Stemming Algorithm (TagSA) [8] to help the process of Text Categorization in detecting words for each category. Stemmer is the process of getting the root word of a given word which includes four (4) steps: Infix Removal, Suffix Removal, Prefix Removal, and Partial or Full Duplication Removal. Every word that was detected will undergo these four processes to stem or trim down the word especially if the detected word is an unknown word.

D. Precision

The act of retrieved instances those are relevant, will also use in the data gathered to measure the degree of its accuracy.

$$\text{precision} = \frac{|\{\text{retrieved documents}\} \cap \{\text{retrieved documents}\}|}{\{\text{retrieved documents}\}} \quad (1)$$

IV. System Architecture

In Fig. 2, it shows that the first process of the system was the tokenization part where the tokens of the sentences with and without speech quotes will undergo a different process. The sentences without speech quoted part will undergo the process of identifying the adjective for a character or the process of gender detection using text categorization. Afterward, it will be sent directly to the text-to-speech part where it will undergo N-gram syllabication and then the output syllables will be used to get its appropriate audio sound in the database of neutral emotion audio files with respect to the detected gender of the character if there is any. Sentences without speech quotes will automatically have a neutral emotion because it was considered as a line of narrator. On the other hand, sentences with speech quoted part will undergo the process of identifying affection word (emotion recognition) in a two-word window. If the system can't find its target, the tokens will be chopped by the stemmer. If an output root word is found and it is an emotional word, transformation-based learning will be applied. The original word will be saved so the system will learn more emotional words. After analyzing the story, the emotion given by the system on each sentence and the gender of the character will be considered wherein the audio sounds will be fetched.

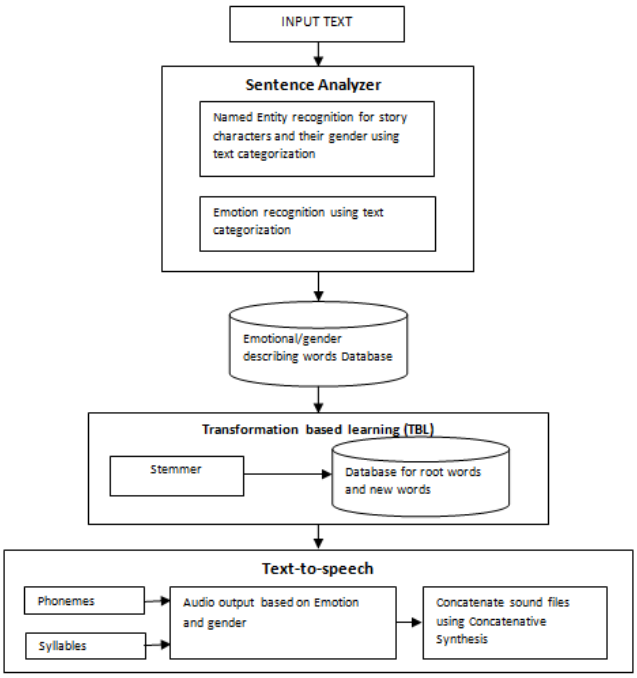

Fig. 2. System architecture of emotional techy basyang.

V. Summary of Findings

TABLE I: Manual and System Count in Emotion Detection

Emotion Detection	Precision	Error Rate
Happy	88.75%	32.57%
Sad	72.23%	16.67%
Fear	87.50%	21.43%
Anger	90.00%	10.00%
Neutral	95.29%	6.70%

Table I illustrates the overall result of the system's accuracy in emotion detection. As what the table implies, the system shows that it is more likely accurate in detecting emotions from text having that the neutral emotion has the most accurate result. The percentage of the precision of the system is higher that its error rate which concludes that it detects emotion precisely.

TABLE II: Manual and System Count in Gender Detection

Gender Detection	Precision	Error Rate
Correct recognition of gender and speech sound of each character	77.38%	15.00%

Table II shows that the system can correctly recognize the gender of the characters in the story and at the same time, it can produce the speech sound of each character based on its detection. As what the result shows, its precision is higher

than its error rate. But a 15% is a little bit high for an error rate.

VI. Conclusion

The precision achieved by the evaluated system are: 88.75% for happy emotion, 72.23% for sad emotion, 87.50% for fear emotion, 90% in anger emotion, and for neutral emotion, 95.29%. When based on its gender detection, the system got a precision of 77.38% which means that the system is almost accurate on detecting gender.

Based on the results, it is concluded that the system can detect emotions accurately. Because the system is more focused on sentiment analysis, we can say that it is possible on detecting emotions from text through sentiment analysis and at the same time, assigning voices to the characters based on its role and emotion throughout the story.

VII. Recommendations

We recommend the enhancement of Named Entity Recognition (NER), Sentiment Analysis, audio used in text-to-speech and audio concatenation.

For the NER, the researchers recommend to detect common nouns as part of character recognition for better detection of characters inside the story. Common nouns are usually used in many stories in describing characters who speak the specified speech quoted sentence/s.

In Sentiment Analysis, include the speech quoted sentence specified in detecting emotion for better detection of emotion or sentiment in every speech quoted sentence.

For Audio concatenation which will be used in Text-to-Speech, the researchers recommended the enhancement of audio recording and digital processing. This will help the audience to understand better the conversion of the text story to speech sounds.

Acknowledgment

Our sincerely gratitude to our families for their love and support during our thesis development, to Ernesto Rondon High School for allowing us to implement in their school and to our Almighty Lord Jesus Christ who gave us the wisdom and knowledge as we finish our thesis papers.

References

[1] H. A. Burrman, *Virtual Storytelling: Emotions for the Narrator*, pp. 7-8, 2007.
[2] K. J. Blom, and S. Beckhaus, "Supporting the creation of dynamic, interactive virtual environments," in *Proc. the 2007 ACM Symposium on Virtual Reality Software and Technology*, pp. 51-57, 2007.
[3] S. Hendrickx, "The virtual storyteller: enriching the story by expressing emotions," *Virtual Storytelling*, vol. 54, 2007.
[4] O. Garin, G. Lobanova, and S. Van Balen, *Fairy Tales for Grownups*, 2005.
[5] D. Doukhan *et al*., *Text and Speech Corpora for Text-to-Speech Synthesis of Tales*, 2011.
[6] Y. S. Seol, H. W. Kim, and D. J. Kim, "Emotion recognition from textual modality using a situational personalized emotion model," *International Journal of Hybrid Technology*, vol. 5, no. 2, pp. 169-174. 2012.
[7] O. Turk *et al.*, "Voice quality interpolation for emotional text-to-speech synthesis," in *Proc. Interspaced 2005*, Lisbon, Portugal, pp. 797-800,2005.
[8] D. E. Bonus, "Tagalog stemming algorithm (Tagsa)," in *Proc. National Natural Language Processing Research Symposium*, Manila, Philippines, pp. 63-67, 2004

John Christopher P. Gonzaga was born on January 15, 1993 in Manila. He received his BS computer science from Polytechnic University of the Philippines Sta. Mesa, Manila.

He took his internshipat PC-Basics Incorporated in San Juan, Philippines as a programmer.

Jemimah A. Seguerra was born on September 6, 1993 in Manila. She received her BS BS computer science from Polytechnic University of the Philippines Sta. Mesa, Manila.

She had her On-the-Job training at the lead companies in Ortigas, Philippines as a website developer for their foundation.

Jhonnel A. Turingan was born on June 9, 1993 in Manila. He received his BS computer science from Polytechnic University of the Philippines Sta. Mesa, Manila.

He had his on-the-job training at Sandiganbayan, Philippines as network administrator and technical support representative at their management information system division.

Mel Patrick A. Ulit was born on February 21, 1994 in Taytay, Rizal. He received his BS computer science from Polytechnic University of the Philippines Sta. Mesa, Manila.

He had his internship at information and communications technology center (ICT center) in Polytechnic University of the Philippines as a website developer for their open university website.

Ria A. Sagum was born in Laguna, Philippines on August 31, 1969. She received her bachelor of computer data processing management from the Polytechnic University of the Philippines and as an professional education at the Eulogio Amang Rodriguez Institute of Science and Technology. She received her master's degree in computer science from the De La Salle University in 2012.

She is currently teaching at the Department of Computer Science, College of Computer and Information Sciences, Polytechnic University of the Philippines in Sta. Mesa, Manila and a lecturer at the information and computer studies, Faculty of Engineering, University of Santo Tomas in Manila.

Ms. Sagum has been a presenter at different conferences, including the 2012 International Conference on e-Commerce, e-Administration, e-Society, e-Education, and e-Technology and National Natural Language Processing Research Symposium. She is a member of different professional associations including ACMCSTA and an active member of the computing society of the Philippines- natural language processing special interest group.