Automatic Speech for Poetry - The Voice behind the Experience

Diana Arellano[1], Cristina Manresa-Yee[2], and Volker Helzle[1]

[1] Filmakademie Baden-Wuerttemberg, Institute of Animation, Germany
[2] Mathematics and Computer Science Dept., University of Balearic Islands, Spain
{diana.arellano,volker.helzle}@filmakademie.de
cristina.manresa@uib.es
http://research.animationsinstitut.de/

Abstract. This paper presents the advances in expressive speech for the interactive installation The Muses of Poetry, as well as the results of an evaluation to assess the interaction and emotional experience of the participants. The objective of the installation is to bring poetry closer to a wider audience through the use of animated characters who not only read poetry, but also manifest the emotional content of the poems. The latter is done using facial expressions and affective speech, which is the focus of this paper. The results of the evaluation show that in general the installation was pleasant and appealing. Also, the three characters managed to convey emotions and awake emotions in the users.

Keywords: Affective Speech, Computational Creativity, User Experience

1 Introduction

Poetry is a form of literary art that can be characterized by its capacity to transport the reader into another, more ethereal world. However, as Kwiatek and Woolner[1] expressed it, *poetry is not always easy to understand, especially for young people.*

In our attempt to bring poetry closer to a wider audience, we developed an interactive installation named The Muses of Poetry, where virtual animated characters recite poetry in an emotional way. A semantic affective analysis of the text of existing poems allows the system to extract their intrinsic affective content, which is manifested by the characters through facial expressions and expressive speech, both automatically generated in real-time.

Unlike previous works that have dealt with poetry and virtual characters [2], automatic generation of poetry with emotional content [3], or without it [4], [5], [6], the work we present combines different fields like real-time computer animation, semantic analysis, human-computer interaction and affective computing, in order to create a believable and engaging expression of the emotions in the poems. We think that The Muses of Poetry might help users not familiar with poetry to understand better the context of the poem, thus awakening an interest

in this art. Moreover, given the real-time feature of our installation, any poem can be analysed, integrated and recited on-site, without the need to spend a large amount of time and resources.

The objective of this paper is twofold. On the one hand, we explain the procedure to enrich the synthetic speech of the characters, or muses with the emotionality of the poems. On the other, we present the results of a user-experience evaluation performed on The Muses of Poetry during a public exhibition.

2 Emotional Speech

One of the characteristics of poetry is its freedom of interpretation, which can lead to different ways of reading it aloud. Pauses, intonation, melody, and emotions are some of the elements that need to be taken into account when reciting poetry. Their correct use can enhance the poetic experience to a level that is capable to engage a wider audience.

In The Muses of Poetry one of our objectives is to transmit the emotionality of the poems not only with visual manifestations, but also with changes in the speech. To that end, a semantic analysis of the text of the poems is performed in order to extract their affective content. From this analysis, both general and "line-per-line", emotional states are obtained, which in the case of The Muses are: pleasant, nice, fun, unpleasant, nasty and sad. As for the poems, these have been provided by real poets from the Cordite Poetry Review Magazine [7], and by poets who have their poems under the Creative Commons licence.

Once the analysis of the text is carried on, the results are integrated into the system, so they can be interpreted by the TTS tool. In our installation we are using the third-party voice synthesizer provided by SVOX[3].

As it is, the TTS produces voices of relative good quality, but with no changes in the prosody. For that reason, we developed a real-time algorithm that indicates, without bias from the human perception, where exactly the changes in the pitch and speed of the voice have to be generated. However, due to the static nature of the poems, this prosodical analysis is performed only once and stored as tags inside the text of the poem.

2.1 Poem into Lines

The structure of the text of the poems follows an XML structure, which facilitates the extraction of the different parts of the poem: author, title and body; as well as the tagging of the emotions.

Before getting into details, it is worth noting that a "line" in the poem is not necessarily the same as the written line (i.e. separated by new lines in the text). We defined a "poetic line" as the set of words, or lines, that enclose one idea of the poem. To divide the text into lines we follow the logic presented in this pseudo-algorithm:

[3] http://www.nuance.de/products/SVOX/index.htm

```
begin
   repeat
     Check if the LINE ends with a period '.'
        if TRUE then
           add it to the list of LINES
           go back to repeat
        else
           if NUMBER_LINES_READ > 3 then
              check for conjunctions in LINE, add a comma before them
              add it to the list of LINES
              go back to repeat
           if LINE ends with a comma ',' then
              go back to repeat
           else if First Letter in NEXT_LINE is capitalized then
              random:
              add it to the list of LINES, OR
              add a long pause '..' at the end of LINE
              go back to repeat
   until endOfFile
end
```

As a result, each time the system encounters a new line or punctuation mark in this new set of lines (`LINES`), it will be interpreted as a pause. The length of the pause depends on the mark, if it is a new line or two points '..', then it is a long pause; on the contrary, if it is a comma ',' or a period '.', then it is a short pause (comma pause is in general shorter than the period pause).

2.2 Tagging the Poem

Once the poem is divided into lines, the semantic affective analysis to extract the emotional states is carried on. The details of the analysis can be found in [8].

As previously mentioned, the result of the global affective analysis of the poem gives one of the following emotional states: pleasant, nice, fun, sad, unpleasant, or nasty. In order to simplify the tagging of the poems, we re-grouped these states in: *happy* (i.e. pleasant, nice, or fun), *unpleasant* (i.e. unpleasant or nasty), and *sad*.

The tagging of a poem is performed line by line, and in the particular case of pitch, also per word. The values we use for speed and pitch are obtained from the global and line-per-line analysis of the poem, according to the following rules:

1. Compute the increment, or decrement, of the neutral pitch and speed, according to the poem emotional state (i.e. if it is happy, sad, or unpleasant). In the case of a human-like character, we use the following values:
   - If the state is *happy*, the neutral speed (i.e. neutral speed=90) is incremented by 5%, and the neutral pitch (i.e. neutral pitch=90) by 10%.
   - For an *unpleasant* state, the decrement in neutral pitch and speed is 8%.
   - For the *sad* state, the decrement in neutral pitch and speed is 5%.

2. Compute for each "poetic line" its resultant emotional value by multiplying (a) the poem emotional state value obtained from the affective analysis using the Whissell Dictionary of Affect in Language [9] (e.g. pleasant = 11.45) by (b) the line emotional state value, which is basically the number of words with the same emotional state of the line (e.g. if the line is pleasant and has 4 words rated as pleasant, then the line state value will be 4)
3. To the speed increased in step 1, add (or subtract, depending on the emotional state) the emotional value obtained in step 2. This value will be used to tag the speed of the whole line.
4. The pitch value to tag the whole line is obtained from step 1. Moreover, for each line, the value obtained in step 2 is added (or subtracted, depending on the emotional state) to the neutral pitch. This value is used to tag the pitch of each word with emotional state similar to the one of the line. If no words are rated with the same state as the line, then no pitch changes are applied.

The subtle variations in step 1 are due to the fact that abrupt changes in a realistic character are seen as very uncanny, diminishing the whole experience. However, this changes in the case of more cartoon or abstract characters, where the common guidelines followed in animation include the idea that cartoon characters should be exaggerated to better convey emotion and intent [10].

The tags we used are provided by the TTS tool, SVOX to change the pitch and speed: `[synthesis:pitch level=PVALUE]`, where `PVALUE` $\in (0, 200)$ and `[synthesis:speed level=SVALUE]`, where `SVALUE` $\in (0, 500)$.

The following excerpt are the "poetic lines" of the poem *For the Road*, by Carol Jenkins. The poem was assessed as *happy* and tagged accordingly:

Line 1

`[synthesis:emotion id='JOY_TRIGGER'][synthesis:pitch level='99'][synthesis:speed level='100']` *First as a dare and then for the* `[synthesis:pitch level= '124']` *warm* `[/synthesis:pitch]` *languor of the tar, at midnight* `[synthesis:pitch level= '124']` *walking* `[/synthesis:pitch]` *to my house, we lay down our bodies on the middle of Moana Road and* `[synthesis:pitch level='124']` *kissed,* `[/synthesis:pitch]` *.. Those long dreamy kisses of abandonment, to each other, to the road, to the dark pines looking on, to the locked light of houses with blinds drawn tight on quarter acre blocks,* `[/synthesis:speed] [/synthesis:pitch ]`

Line 2

`[synthesis:emotion id='PLEASANT_TRIGGER'][synthesis:pitch level='99'] [synthesis:speed level='100']` *The stars' bright and dizzy mass arcing over us, and we'd get to our feet,* `[synthesis:pitch level='101']` *like* `[/synthesis:pitch]` *angels coming to in a strange world,* `[/synthesis:speed] [/synthesis:pitch]`

Line 3

`[synthesis:emotion id='JOY_TRIGGER'][synthesis:pitch level='99'][synthesis:speed level='100']` *To walk down the road, arms and hands tangling,* `[synthesis:pitch level='124']` *laughing, like* `[/synthesis:pitch]` *we'd swallowed a* `[synthesis:pitch level='124']` *universe* `[/synthesis:pitch]` *and it was exploding out of our fingertips.* `[/synthesis:speed] [/synthesis:pitch]`

It can be seen that at the beginning of each line, the pitch and speed have the same value (pitch = 99, speed = 100). Nonetheless, the pitch-tagged words (e.g. *warm*, *walking*, and *kissed*) have different values (*Line 1 y 3*, pitch = 124; *Line 2*, pitch = 101), obtained from the rules previously explained.

3 User Experience

To assess the interaction and the emotional engagement produced by The Muses of Poetry, we carried on an user evaluation during its exhibition at FMX 2013, the Conference on Animation, Effects, Games and Transmedia. This conference reunites professionals, students and amateurs focused in the development, production and distribution of digital media, as well as computer graphics and animation, all of whom constituted the main audience who interacted with The Muses of Poetry.

Regarding the installation, it was a stand resembling an open book, where the slides that simulated the pages formed a kind of cave, where the user could enter and interact with the characters. The interaction was thought to be free-of-devices, for which we installed a clip microphone on the top of the second slice, almost invisible to the human eye, and approximately 1 meter away from the projection screen.

As for the virtual characters, we designed and implemented three types that could cover the wide spectrum of animated characters. The first one was a realistic human-like female, Nikita, which we have used in previous research applications. In this case, we added a veil to make her look more ancient-Greek like. The second was designed by one of the students at the Institute of Animation, following the premise of a more abstract character. This was made of particles that are shaped in the form of a human head, and were constantly moving when the character spoke. The third one, Myself, was a 2D cartoon character designed by the German animator Andreas Hykade, which came to complement this first repertoire of muses. Figure 1 shows the installation and the evaluated characters.

Participants experienced the installation alone with the support of a member of the development team, who explained how the system worked. Each participant interacted with only one character: Nikita, Particles, or Myself. At the beginning, the character asked the user to select two words from a word cloud displayed on the screen. Once the two words were recognized, a poem containing those two words was selected and recited. At the end of the experience, the participant completed a questionnaire regarding the installation, the character and his or her feelings towards the interactive system and the poem reading. The written questionnaire consisted in eight 5-point Likert scale questions ranging from “Strongly Disagree” on one end to “Strongly Agree” on the other:

Q1: The character is attractive as a poetry reader
Q2: The character conveys emotion
Q3: The character read the poem in a way I understood the topic of the poem
Q4: I think I would like to visit this installation frequently or I would recommend it to my friends

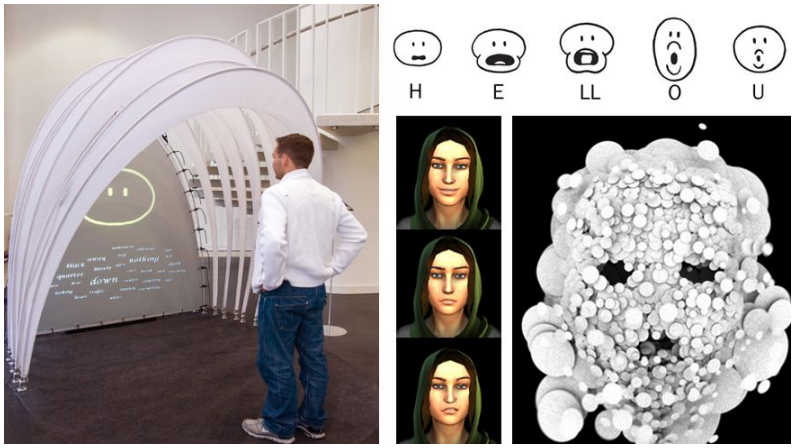

Fig. 1. Left: Installation and location of the mic. Right: (top) Myself, 2D character; (down-left) Nikita, 3D realistic character; (down-right) Particles, 3D abstract character

Q5: I felt a variety of emotions while listening to the poems
Q6: The system is pleasant
Q7: The system is inviting
Q8: The system is appealing

In the end 51 questionnaires were gathered: 35 from male participants and 16 from female participants, with ages ranging between 19 and 45 years (average age was 27). Data was analysed for each specific question using two approaches: separately for each character to find differences among the three poetry readers and together to conclude global insights from the interactive system.

Figure 2 shows the boxplots for each of the eight questions, considering the results of the three characters together. Based on the interquartile ranges (IQR) of questions Q1, Q2 and Q5, all related with the emotional aspect of the installation, we could not conclude if the user felt emotions when listening to the poems, or if he or she felt the characters emotional enough, because the bars are widely distributed. However, Q3, which evaluated the degree of empathy with poetry, throws *median*=4, which indicates that one of the objectives of the installation (make a wider audience understand poetry) might have been achieved. As for questions Q6, Q7 and Q8, despite the outliers, we can conclude that the installation was pleasant, inviting and appealing.

Figure 3 shows the boxplots for each of the eight questions, for each character. The boxplot on the right of figure 3 shows the IRQs when evaluating Nikita. For Q1, although the median value was above the mid point, it cannot be concluded that Nikita was attractive as a poet reader. The analysis of Q2 and Q3 do not throw decisive results either. However, people felt they understood the topic of the poem (Q4) and felt the installation more inviting (Q7) and appealing (Q8) than pleasant (Q9). From the boxplot corresponding to Particles we can conclude

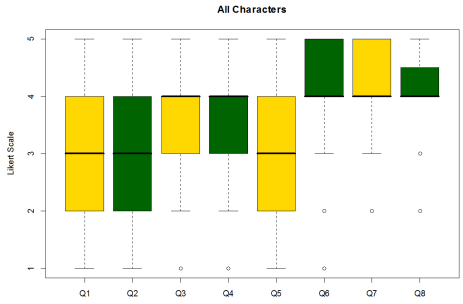

Fig. 2. Inter-quartile analysis for each question and all characters

that people indeed understood the content of the poems (*median*=4), although 50% of the sample is under this value. As for the appealing, pleasantness and inviting nature of the installation, results are not so satisfactory as with Nikita and Myself. Nonetheless, these elements were rated as positive. Finally, from the boxplot with the evaluation of Myself, we cannot conclude that the character was attractive as a poet reader (Q1), but it did conveyed emotions (Q2) and 50% of the sample felt emotions while listening to the poems. Participants also felt that with this character the installation was pleasant, inviting and appealing. A last thing to note is that with Myself users would recommend the installation to friends, in a higher scale than with the other characters.

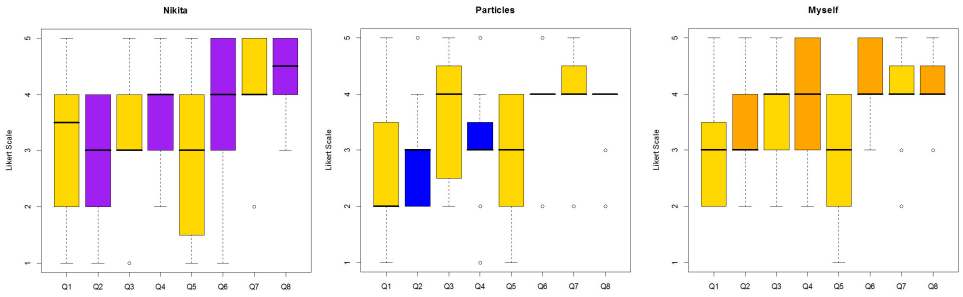

Fig. 3. Inter-quartile analysis for each question and each character: Nikita (left), Particles (middle), Myself (right)

A last analysis focused on the mode and median differences based on genders showed that women found the interactive installation more inviting and pleasant than men, and in general they rated higher values than men in all question, as seen in Table 1. In a more detailed analysis, women also felt more likeliness for the Myself character, while men for Nikita. Regarding the mode, it was observed that women felt more emotions while listening to the poem.

Table 1. Data of participants and evaluated characters

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**
median (women)	3	3	4	4	3	4.5	5	4
median (men)	3	3	3	4	3	4	4	4
mode (women)	3	3	4	5	4	5	5	4
moda (men)	2	3	3	4	2	4	4	4

4 Discussion and Future Work

We presented the advances in affective speech generation and interaction results of the on-going project The Muses of Poetry, an interactive installation where animated characters recite poetry in an emotional way. Regarding speech, the generated affective voices were in general satisfactory, enhancing the conveyance of emotions. The results of the user experience evaluation showed that the majority of the participants found the installation pleasant, inviting and appealing. However, these results did not allow us to conclude which character performed better as a poetry reader. In the future we will continue working on the speech generation to produce a more natural intonation, we will add more emotional expressions in the current characters, as well as new animated characters.

Acknowledgments. The project is funded by the *Innovationsfonds Kunst* of the Ministry for Science, Research and Art in Baden-Württemberg (54-7902.513/65)

References

1. Kwiatek, K., Woolner, M:. Let me understand the poetry. Embedding interactive storytelling within panoramic virtual environments. In: EVA 2010, pp. 199–205 (2010)
2. Tosa, N.: Interactive poem. In: ACM SIGGRAPH 98 Conference abstracts and applications, SIGGRAPH 98, pp. 300 (1998)
3. Kirke, A., Miranda, E.: Emotional and Multi-agent Systems in Computer-aided Writing and Poetry. In: Symposium on Artificial Intelligence and Poetry, pp. 17–22 (2013)
4. Colton, S., Goodwin, J., Veale, T.: Full-FACE Poetry Generation. In: Proceedings of the 3rd International Conference on Computational Creativity, pp.95–102 (2012)
5. Gervás, P., Hervás, R., Robinson, J. R.: Difficulties and Challenges in Automatic Poem Generation: Five Years of Research at UCM. In: E-Poetry 2007 (2007)
6. Cope, D.: Comes the Fiery Night. NY: Amazon Books (2011)
7. Cordite Poetry Review Magazine, `http://cordite.org.au/`
8. Arellano, D., Spielmann, S., Helzle, V.: The Muses of Poetry - In search of the poetic experience. In: Symposium on Artificial Intelligence and Poetry, pp. 6–10 (2013)
9. Duhamel, P., Whissell, C.: The dictionary of affect in language [software] (1998)
10. Hyde, J., Carter, E. J., Kiesler, S., Hodgins, J. K.: Perceptual effects of damped and exaggerated facial motion in animated characters. IEEE International Conference on Automatic Face and Gesture Recognition - FG 13 (2013)