An Affective Spoken Storyteller

Felix Burkhardt

Deutsche Telekom Laboratories, Berlin, Germany

Felix.Burkhardt@telekom.de

Abstract

We present a software to read texts with emotional expression. The software is developed as part of the Emofilt open source emotional speech synthesis software. The affective storyteller consists of a text editor which offers a set of emotional speaking styles that can be used to mark up the text. The system was validated in a perception experiment and, although the number of participants wasn't very large, could show the general usability of the approach.

Index Terms: speech synthesis, emotional, application

1. Introduction

Human Machine Interaction becomes more and more part of our daily life, emotional processing can enhance entertainment as well as communication systems. While applications for emotion recognition, like stress detection, come easy to mind, the automatic expression of emotions by computer systems, although researched since approximately two decades now, has not yet led to many practical applications. One application is the speech synthesis of stories that include characters displaying personalities and emotional states.

We developed a software to read texts with emotional expression and evaluated its usefulness within a perception experiment. The software is developed as part of the Emofilt open source emotional speech synthesis software [1] Emofilt as well as the affective storyteller is freely available[1] and the reader is invited to reproduce the simulations. There is a lot of research on emotional speech synthesis (for a review, see [2]), and also emotional story tagging approaches have frequently been described in the literature. The ESPER project [3] is concerned with the identification of quotes and assignment to speaker characters in children's stories but does not itself generates emotional speech. It would thus make a good combination with the affective storyteller by proposing automatically quotes and character names to the system. In [4], the sequencing of fairy tales sentences, annotated with eight basic emotions, was investigated. [5] developed something similar to the application described in this paper, the "Manual Emotion Annotation Tool" can be used to annotate text sections emotionally.

The affective storyteller consists of a text editor which offers a set of emotional speaking styles that can be used to mark up the text. Section 2 describes Emofilt, the system that is used to define the acoustic properties of an emotional speaking style. After this, in section 3 the affective storyteller interface gets described, followed by an evaluation experiment in section 4. Section 5 concludes the paper.

[1] http://emofilt.sourceforge.net

2. Emofilt

The emotional simulation is achieved by a set of parametrized rules that describe manipulation of certain acoustic aspects of a speech signal, which is done by Emofilt [1]. Emofilt is an open source program to simulate emotional arousal with speech synthesis based on the free-for-non-commercial-use MBROLA synthesis engine [6] that is available for over 44 languages. A set of rules can be stored under a specific name as an emotional speaking style and then be used by the affective storyteller.

The rules were motivated by descriptions of emotional speech found in the literature in addition to some speaking styles appropriate for fairy tale story telling that were adjusted manually.

Before the rules are applied by Emofilt, the input phoneme chain gets syllabelized by an algorithm based on sonority hierarchy. In addition, stressed syllables are identified as those that carry local pitch contour maxima.

The following paragraphs describe possible modifications provided by Emofilt.

The pitch contour of the whole input as well as selected syllables can be modified by altering either the level, the range or the form of contour. The speech rate can be modified for the whole phrase, specific sound categories or syllable stress-types separately by changing the duration of the phonemes (given as a percentile).

Since with Mbrola the voice quality of the speech is fixed within the diphone inventory, we had to restrict ourselves to Jitter and vocal effort in terms of voice quality modification. In order to simulate jitter (fast fluctuations of the F_0-contour) the F_0 values can be displaced by a percentile alternating down and up. With respect to Vocal effort; for the German language exist two voice-databases that were recorded in three voice-qualities: normal, soft, and loud [7].

Considering articulation modification, a diphone synthesizer has a very limited set of phoneme realizations and does not provide for a way to do manipulations with respect to the articulatory effort. Thus, the substitution of centralized vowels with their decentralized counterparts and vice versa is possible as a work-around to change the *vowel precision*.

3. The affective storyteller

The affective storyteller is an extension of the Emofilt software and consists of an editor in which parts of text can be marked. The graphical user interface is depicted in Figure 2. The set of emotional speaking styles that was specified by Emofilt is automatically imported into the GUI and can be used to mark up arbitrary parts of the text. The application inserts a newline character automatically after each style change, which is additionally signaled by a text color unique for each emotional style. Texts can be loaded, saved and synthesized either phrase by phrase or as a whole and saved as an audio Wave file.

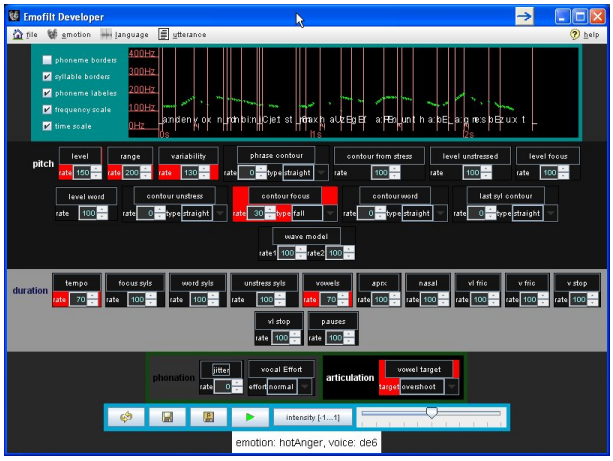

Figure 1: *The Emofilt graphical user interface.*

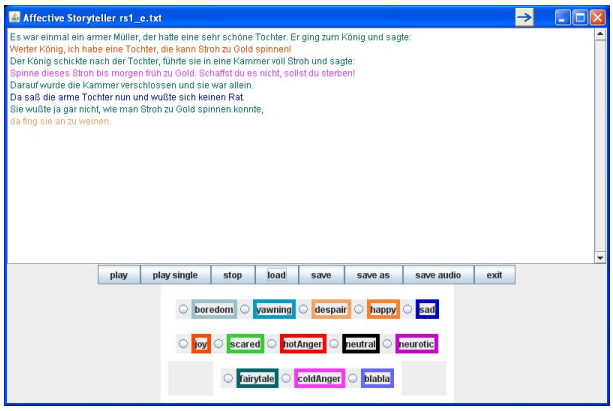

Figure 2: *The affective storyteller graphical interface.*

4. Evaluation

We evaluated the application within a perception experiment. 15 children aged 6 to 16 (mean 12.6 years) listened to different versions of the German fairy tale ”Rumpelstilzchen”. The story was shortened and manually annotated with the affective storyteller. The assigned speaking styles were chosen to be appropriate for the personality of the character as well as content of the quote. The following emotion or personality styles were derived from literature descriptions and adjusted manually: ”bored”, ”despaired”, ”happy”, ”sad”, ”joyful”, ”scared”, ”neurotic”, ”fairy tale”, ”cold anger” and ”hot anger”.

The annotated text, as well as a non-emotional version for comparison, was synthesized with the affective storyteller with one of the German MBROLA voices (de6). Each version lasts about five minutes. Eight children listened to the non emotional version and seven to the version synthesized with the affective storyteller. Afterwards they were asked for a short questionnaire containing the following questions.

- how they liked the story (in German school grades)
- how they liked the speaking style (in German school grades)
- how many facts (max. 7 detail facts) they remembered from the story

The results are displayed in Figure 3. Box-plots are shown

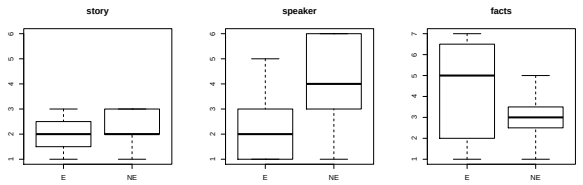

Figure 3: *Listener judgments for liking of story, speaker style and number of remembered facts when listening to the emotional (E) and non-emotional (NE) version of the fairy tale.*

for the listener judgments for liking of story, speaker style and number of remembered facts when listening to the emotional (E) and non-emotional (NE) version of the fairy tale. We computed an analysis of variance on these results. Only the factor ”speaking style” results in a a strong tendency for a significant difference of judgment for the two versions (level of significance .0589). We realize though, that the number of participants for this test was too low to make resilient predictions. Although some children criticized the lack of understandability resulting from the emotional way of speaking, they did like the resulting variation.

5. Conclusions and outlook

We described a system for emotional tagging of text parts and subsequent synthesis with an appropriate speaking style. The system was validated in a perception experiment and, although the number of participants wasn't large, could show the general usability of the approach. In order to further automate the process. the system could be connected to systems that identify speakers automatically like [3] and/or statistical sentiment analysis systems.

Another aspect that turned up in scope of this research study is the question on how to overlay the characters personalities, that stay constant during the story and are important to recognize the character, with specific emotions that are displayed during one quote.

6. References

[1] F. Burkhardt, “Emofilt: The simulation of emotional speech by prosody transformation,” in *Proc. Interspeech 2005, Lisbon*, 2005.

[2] M. Schröder, “Emotional speech synthesis - a review,” in *Proc. Eurospeech 2001, Aalborg*, 2001, pp. 561–564.

[3] J. Y. Zhang, A. W. Black, and R. Sproat, “Identifying speakers in childrens stories for speech synthesis,” *Proc. Eurospeech, Geneva*, 2003.

[4] C. Alm and R. Sprout, “Emotional sequencing and development in fairy tales,” *Proc. ACII*, 2005.

[5] E. P. Volkova, B. Mohler, D. Meurers, D. Gerdemann, and H. Bülthoff, “Emotional perception of fairy tales: Achieving agreement in emotion annotation of text,” *Proc. of Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, NAACL HLT*, 2010.

[6] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vreken, “The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes,” *Proc. ICSLP'96, Philadelphia*, vol. 3, pp. 1393–1396, 1996.

[7] M. Schröder and M. Grice, “Expressing vocal effort in concatenative synthesis,” in *Proc. 15th International Conference of Phonetic Sciences, Barcelona, Spain*, 2003, pp. 2589–2592.