Generating Expressive Speech for Storytelling Applications

Mariët Theune, Koen Meijs, Dirk Heylen, and Roeland Ordeman

Abstract—Work on expressive speech synthesis has long focused on the expression of basic emotions. In recent years, however, interest in other expressive styles has been increasing. The research presented in this paper aims at the generation of a storytelling speaking style, which is suitable for storytelling applications and more in general, for applications aimed at children. Based on an analysis of human storytellers’ speech, we designed and implemented a set of prosodic rules for converting ‘neutral’ speech, as produced by a text-to-speech system, into storytelling speech. A small-scale evaluation of our storytelling speech generation module showed encouraging results.

Index Terms—child-directed speech, expressive prosody, expressive speech, speech analysis, speech synthesis

I. INTRODUCTION

So far, most research on expressive speech synthesis has been aimed at the prosodic expression of ‘basic’ emotions such as sadness, fear, happiness and anger. However, many speech applications require other expressive speaking styles in addition to, or instead of, the expression of emotions. Here we focus on one particular speaking style: storytelling speech.

Human storytellers use their voice in a variety of ways to capture their audience’s attention. They mimic characters’ voices, produce various ‘sound effects’ and use prosody to convey and invoke emotions, thus creating an engaging listening experience. In digital storytelling, stories are told by a computer. Ideally, a digital storytelling application should deliver a listening experience that is equally engaging as that provided by a human storyteller. To achieve this ideal we need a far more expressive, engaging speaking style than is provided by today’s text-to-speech systems. In this paper we describe a first step in this direction: the development of a software module for the automatic generation of speech with storytelling prosody.

The context of our work is the Virtual Storyteller, a story generation system developed at the University of Twente [1]. In this system, story plots are automatically created based on the actions of intelligent agents living in a virtual story world. The generated plots are converted to natural language, and presented to the user by an embodied agent that makes use of text-to-speech. Currently this is done in a crude fashion, using fixed templates for language generation in combination with a standard (commercial) text-to-speech system. However, we are currently working on improving these aspects of our system. In this paper we focus on our improvements to the speech output of the Virtual Storyteller. To make the speech produced by the system more suitable for storytelling, we have focused on two aspects: first, the creation of a general storytelling speech style, and second, the prosodic expression of suspense. Our general approach has been to perform an analysis of storytelling speech and based on this, design a set of rules that modifies the prosodic parameters produced by a standard Dutch text-to-speech synthesis system called Fluency (http://www.fluency.nl). Fluency is a commercial diphone synthesis system which does not allow us to modify voice quality. For that reason our work is restricted to prosody, even though we expect that differences in voice quality also play an important role in human storytelling speech.

The paper is structured as follows. In Section II we give an overview of related work. In Section III we describe the main patterns we observed in human storytelling speech, focusing on global speech style and the expression of suspense. We describe our rules for converting neutral speech to storytelling speech in Section IV, and the implementation of these rules in Section V. In Section VI we describe the evaluation of our storytelling speech generation module. We end with a discussion and conclusions in Sections VII and VIII.

II. RELATED WORK

The first attempts to improve speech synthesis by adding human-like expressivity have focused on the expression of emotions. The earliest, rule-based systems for emotional speech generation are the Affect Editor [2] and HAMLET [3]. More recent, concatenative approaches include that of [4], who synthesized four basic emotions using a combination of prosodic rules and specific diphone inventories for each emotion, and [5] who used a unit selection approach to generate a happy ‘Genki’ speech style. Recent approaches to emotional speech generation in languages other than English include [6] for Dutch, [7] for Spanish, [8] for Catalan, and [9] for German. All the latter approaches are rule-based, like ours. A distinguishing feature of [9] is the focus on emotion dimensions rather than on a small set of ‘basic’ emotions.

In recent years, interest in non-emotional expressive speaking styles has been growing. It has been recognized that depending on the domain and the target group of speech applications, different expressive styles are required. For example, in applications aimed at children, highly expressive speech has been shown to greatly increase the ‘fun’ factor.

M. Theune, D. Heylen and R. Ordelman are with the Human Media Interaction group at the University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. They participate in the EU Network of Excellence HUMAINE. E-mail: {m.theune | d.k.j.heylen | r.j.f.ordelman}@utwente.nl

M. Theune is supported by NWO (Netherlands Organization for Scientific Research), grant number 532.001.301.

E-mail Koen Meijs: koen@nephila.nl
Fig. 1. Newsreader intensity and pitch. (Translation: ‘Around 50 municipalities have participated in the carless Sunday.’)

Fig. 2. Children’s storyteller intensity and pitch. (Translation: ‘that Jelmar had to be careful not to bump his head.’)

[10]. For a training application in the military domain, [11] used a limited domain unit selection approach to generate shouted commands, spoken commands, shouted conversation and spoken conversation. To generate speaking styles for different dialogue contexts, [12] trained prosody models for expressing good news, bad news, and questions; as part of a pilot study into expressive speech synthesis, [13] have successfully built small unit selection databases for expressive, sad and angry speaking styles. A prosodic analysis of the reading styles for different text types such as stories, news, and technical documents is provided by [14].

An embodied digital storyteller that can express emotions through prosody and facial expressions has been developed by [15]. To achieve emotional speech output, they adapt the prosodic parameters of a text-to-speech system based on tags in the text of the story [16]. The importance of expressing suspense in storytelling is pointed out by [17]. They adapt the expressivity of their storyteller depending on ‘suspense progression’, ‘narrative conflict’ and ‘narrative relevance’ of the different scenes in the story. However, in their system the different levels of expressivity are only reflected by facial expressions and gestures, not by prosody.

III. HUMAN STORYTELLING SPEECH

As a first step in our research we performed an informal analysis of the speech of a few human storytellers. Our target application, the Virtual Storyteller, generates fairy tales. Therefore, as the main material for our analysis we used existing, commercial recordings of children’s fairy tales told by professional Dutch voice actors. We randomly selected five stories of 5 to 12 minutes long, with three different male storytellers. As comparison material we used recordings of four short (30 seconds to 1.5 minute) radio news broadcasts. Comparing the storytellers’ speech style with the more neutral speech of the (male) newsreaders gave us an idea of the main prosodic differences between the two speaking styles. Based on this, we decided which kind of changes should be made to the neutral speech generated by a standard text-to-speech system to make it more suitable for storytelling.\(^1\) Below we describe our general observations of our speech material, focusing on the storytellers’ global speech style and the way they used prosody to express suspense.

A. Global Speaking Style

When informally comparing our storyteller and newsreader speech material, we observed that the storytellers’ speech showed much more variation in pitch and intensity than the newsreaders’ speech. This is illustrated in Fig. 1 and 2, showing the intensity and pitch contours of a newsreader and a storyteller speech fragment. (Grey line is intensity, black line is pitch.) Like [14], we found that the storytellers tended to speak slower than the newsreaders, and take longer pauses, in particular between sentences. Finally, the storytellers often added extra emphasis to certain adjectives and adverbs by increasing their pitch and duration. For example, ‘A loooong corridor ...’ This ‘vowel stretching’ typically occurred with words indicating an extreme value of some property.

B. Expressing Suspense

A storyteller’s main goal is to capture the audience’s attention and keep them engaged with the story. One way of doing

\(^1\)This approach presupposes that the default speech synthesis is roughly equivalent to human newsreading style. We did not check this assumption.
that is by using prosody to build suspense. Using only his voice, the storyteller can create a feeling of expectation, and warn the audience that something exciting is about to happen.

Two kinds of suspense, as signaled by the storytellers’ prosody, could be distinguished in our material, with each of the stories containing several examples of each kind. The first type was the sudden climax: an unexpected dramatic moment in the story, such as a startling revelation or a sudden, momentous event. For example, ‘Everything was peaceful, but suddenly there was a loud noise.’ Typically, in our material such climactic events were announced by a steep increase of intensity and pitch on the keyword introducing the climax (‘then’, ‘suddenly’, ‘but’...). This is illustrated in Fig. 3, which shows the intensity and pitch contour for ‘and then’ in a fragment from the story of Bluebeard, where the unsuspecting princess opens the door to Bluebeard’s secret chamber. (‘Her eyes had to get used to the darkness, and then ...!’)

The second type of suspense is the increasing climax, where the dramatic event is expected in advance. We observed that when approaching the climax, our storytellers heightened the suspense by a gradual increase in pitch and intensity, accompanied by a decrease in tempo. They typically added a pause before the description of the actual dramatic event. Thus, by postponing the revelation, they built up dramatic tension. This is an example of an increasing climax from Sleeping Beauty: after the prince has slowly made his way through the thorns to Sleeping Beauty’s chamber, ‘He opened the door and... there was the sleeping princess.’ Fig. 4 shows the intensity and pitch contour of this fragment, clearly showing the pause between en (‘and’) and daar (‘there’). We also see a decrease in pitch, intensity and duration after the pause.

IV. FORMULATING STORYTELLING PROSODY RULES

In this section we describe the rules we designed for converting neutral speech to storyteller style speech, and for creating suspense effects as outlined in Section III-B. The rules take as input a list of paired time-value data representing the neutral prosodic features of a given utterance, and return new values for these features. First, we describe our rule design method, then we give the most important rules involved in creating a global storytelling style and the two types of suspense we distinguish.

A. Method

Our approach in creating the storytelling prosody rules was as follows. First we made a global rule design, based on our observations described in the previous section. Then some representative samples from our speech material were analyzed in detail, using the speech analysis tool Praat [18]. Based on this analysis, we determined the possible range of certain constants to be used in our prosodic rules (e.g., pitch and intensity increase with respect to neutral speech).

To determine the best values within this range, we performed a small perception test. In this test, five subjects were presented with 23 pairs of synthesized text fragments. For each pair, they had to indicate which version they found the most natural sounding (for general storytelling speech style) or the most suspenseful (for increasing or sudden climax).

The utterance pairs used in the test were created as follows. First, we synthesized some sentences from our fairytale stories using the Fluency Dutch text-to-speech system. Five of these text fragments expressed a sudden climax, six fragments expressed an increasing climax, and the rest were neutral (i.e., non-suspenseful). Then, we manipulated the resulting speech using Praat scripts [18]. These scripts simulated the prosody rules, with different values for the constants we wanted to test. For each pair of utterances in our test, the two versions only differed from each other with respect to the value of one constant. We also included some sentences with fully neutral prosody. The samples we took from the storyteller speech material were neutral in the sense that they did not contain any ‘special effects’ such as emotional speech or mimicking the voice of a story character. We took care to avoid such effects, because they would have influenced the analysis of
the aspects we were really interested in, i.e., global storytelling style and the expression of suspense. We ignored voice quality differences such as the use of creaky or breathy voice, because we would not be able replicate these using Fluency.

Below we discuss the prosodic rules we designed. For each rule we present the range of constants we found in our human speech data, the range of constants we actually tested, and the outcome of the perception test.

B. Rules for Global Storytelling Speaking Style

The following rules can, in combination, be used to change a neutral speaking style into a general storytelling speech style by modifying the prosodic parameters pitch (only of accented syllables), intensity (only of accented syllables), overall speech tempo (in syllables per second), overall pause duration, and vowel duration in certain adjectives and adverbs.

For pitch, we found in our analysis that storytellers use relatively larger pitch excursions than newscasters, so we designed a rule that manipulates the pitch contour of the accented syllables in words carrying sentence accent. The rule multiplies all pitch values within the relevant time domain \([t_1, t_2]\) by a factor based on (the first part of) a sine wave form. The sine function is used to ensure that the pitch is increased gradually within the given time domain. This is a crude approximation of the up-down pitch contour which is increased gradually within the given time domain. (See [19] for an analysis of storytelling pitch contours.) The rule looks as follows:

\[
y'(t) = y(t)(1 + a * \sin(\frac{t - t_1}{t_2 - t_1} - 0.5\pi + 0.25\pi))
\]  

where

\(t \in [t_1, t_2]\)

\(y\) is the original pitch value \(y\) as a function of \(t\)

\(y'\) is the modified pitch value

\(a\) is desired maximum pitch increase divided by average pitch

To determine the desired maximum pitch increase, we analyzed three storytelling fragments (total length 17.4 seconds). On the syllables carrying a sentence accent we found pitch increases between 40-90 Hz relative to the speaker’s average pitch. In our perception test we slightly decreased the lower bound of this range, because we suspected that lower values might sound more natural. Also, fragments with an increase higher than 60 Hz sounded too unnatural to even include in the evaluation. Therefore we tested a range of pitch increases between 30-60 Hz. The best value we found was 40 Hz.

Because accented syllables in storyteller speech tend to have a relatively high intensity, we also designed a rule for intensity increase. Within the time domain \([t_1, t_2]\) of a syllable carrying sentence accent, this rule simply increases the intensity with a constant value. In our storytelling speech material (the same fragments as used for pitch analysis), we observed increases between 4-7 dB relative to the speaker’s average intensity. For the same reasons as mentioned above, we tested a lower range (2-6 dB) in our experiment, and found 2 dB as the best value.

Storytellers’ speech typically has a slower tempo than neutral speech. We analyzed a few storytelling and newsreading fragments of about five sentences each, and found average speaking rates of 3.0-3.6 syllables per second for storytelling, against 5.8 syllables per second for news reading. XXXX In our perception test, a rate of 3.6 syllables per second was found to be most natural for storytelling, so we used that for the general storytelling speaking style. XXXX HAS THIS REALLY BEEN TESTED? XXXX

In addition to the general slowdown in tempo, in our analysis we found a duration increase in the accented vowel of certain adjectives and adverbs, typically indicating some extreme value of a property. For two of these words we measured the duration of the accented vowel and found that it was stretched to 1.4 and 1.8 times its average duration respectively. Our perception test only included one fragment in which vowel stretching could be applied (cf. the example in Section III-A), and for this fragment the subjects indeed preferred the version with a 50% vowel duration increase over the version without increase in vowel duration.

Finally, by slowing down the overall tempo we already lengthen the pauses in the speech, but the differences in pause length between storytellers and newscasters are larger than predicted by the global difference in tempo. We therefore added a rule fixing pause length at 0.4 seconds at phrase breaks within sentences, and at 1.3 seconds between sentences. These values were based on the average lengths of the phrase and sentence breaks found in the fragments that were used to determine average speech tempo. These values were not tested in the perception experiment. For comparison: in the newscaster speech material we found average pauses of 0.3 and 0.5 seconds respectively.

C. Rules for Sudden Climax

Below we discuss the rules used to express a sudden climax. The time domain \([t_1, t_2]\) referred to in these rules is that of the accented syllable in the word announcing the climax (e.g., ‘suddenly’, ‘then’). Within this time domain, pitch, intensity and vowel duration are strongly increased. The value ranges mentioned below were found by analyzing two speech fragments expressing a sudden climax (the one shown in Fig. 3 and another, similar fragment).

For a sudden climax, the increase in pitch is abrupt and constant throughout the target time domain, i.e., the keyword announcing the climax. The fragments we analysed had pitch increases of 80 and 120 Hz respectively. In our perception test we tested both, and the best value turned out to be 80 Hz.

In a sudden climax, intensity is abruptly increased at \(t_1\) but then gradually decreases to its normal value at \(t_2\). In our speech fragments, we observed initial intensity increases of 6 and 10 dB relative to the speaker’s average intensity. The best value for \(c_{sc}\) found in our test was 6 dB.

Finally, as in the rules for global storytelling style, we applied a duration increase of 50% to the vowel of the word announcing the sudden climax. This value was not tested.

D. Rules for Increasing Climax

The time domain for the increasing climax is split up into two parts, both typically spanning a clause. The first part builds up the expectation and ends with the key word announcing the
where each accented syllable within time domain \([t_1, t_2]\) in the first part of an increasing climax equals average pitch divided by desired maximum pitch increase. The original pitch value \(y\) as a function of \(t\):

\[ y(t) = y(\cdot)(1 + a_{ic} \cdot \sin(\frac{t - s_i}{s_j - s_i} \cdot 0.5\pi + 0.25\pi)) \]  

where

- \(t \in [s_i, s_j]\)
- \(y\) is the original pitch value \(y\) as a function of \(t\)
- \(y'\) is the modified pitch value
- \(a_{ic}\) is a constant specifying the size of the pitch adjustment

\(=\) average pitch divided by desired maximum pitch increase in the first part of an increasing climax.

Because we want to gradually enlarge the pitch increase of each accented syllable within time domain \([t_1, t_2]\), in \(2\) the desired maximum pitch increase is not a constant value, but depends on the position of the syllable within the time domain. Based on the start time \((s_i)\) of the syllable, relative to \([t_1, t_2]\), we compute which fraction of the total pitch increase between \(t_1\) and \(t_2\) should be applied to the syllable:

desired max. pitch increase = \(p_1 + (p_2 - p_1) \cdot \frac{s_i - t_1}{t_2 - t_1}\)  

where

- \(p_1\) is a constant specifying the desired pitch increase at \(t_1\)
- \(p_2\) is a constant specifying the desired pitch increase at \(t_2\)

For the increasing climax, we only analysed one storytelling speech fragment in detail (the one shown in Fig. 4). Here we found a pitch increase of 100 Hz at \(t_1\) (the start of the increasing climax) and of 130 Hz at \(t_2\) (the ‘top’ of the climax). Since these values did not sound acceptable when we reproduced them for our experiment, we decided to shift down the value ranges to be tested for \(p_1\) to 25-50 Hz and for \(p_2\) to 60-80 Hz. The perception test resulted in 25 Hz and 60 Hz as the best values for \(p_1\) and \(p_2\). This means that at time \(t_1\) we apply an initial pitch increase of about 25 Hz, and from \(t_1\) to \(t_2\) we gradually increase this to 60 Hz. As specified in \(2\), this increase is applied only to the accented syllables to be found within \([t_1, t_2]\).

In addition to the pitch increase between \(t_1\) and \(t_2\), we apply an intensity increase of 10 dB. This increase is constant across \([t_1, t_2]\). Finally, between \(t_1\) and \(t_2\), there is a gradual increase in the duration of accented vowels, towards a maximal duration increase of 150% at \(t_2\). At \(t_2\) our rules insert a 1.04 sec pause. This is just before the revelation of the climactic event that the first part of the increasing climax has been leading up to. The actual description of this event takes place in the second part of the increasing climax, which has time domain \([t_2, t_3]\).

In \([t_2, t_3]\), pitch gradually decreases to its normal value. This is done using rules analogous to \(2\) and \(3\). Vowel duration is also gradually decreased to its normal value. For intensity, we see a different pattern: there is a 6 dB intensity increase on the first accented syllable in \([t_2, t_3]\), and after that the rest of \([t_2, t_3]\) is spoken with average intensity without need for modification. Except for the pitch values used in Rule 3, all constants used in the increasing climax rules were based on the values found in the single analysed speech fragment, and were not tested in the perception experiment.

### E. Discussion

If we look at the outcomes of the perception test, which we carried out to determine the best values for most of the constants used in our rules (see Section IV-A), an interesting observation can be made: almost unanimously, the subjects preferred the values near the lower bound of the value ranges of the different constants. According to the comments made by the participants, higher values were perceived as ‘unnatural’ or ‘too much’. Apparently, the prosodic extremes that naturally occur in human speech are less acceptable in the context of synthetic speech. Another possible explanation is that the more extreme modifications, using the maximum values of the constants used in our rules, caused unexpected artifacts in the speech that was generated.

The constant values we tested in the experiment, and which we use in our rules, were based on the analysis of a very small number of speech fragments. Not all constant values we use in the rules were tested in the perception experiment, and for those we tested it is very well possible that the optimal values lie outside the tested range, closer to the default value generated by the text-to-speech system. This means that the concrete values we use in our rules should be regarded as highly preliminary; most likely, these values are not optimal. On the other hand, they do seem to be a step in the right direction, as some participants spontaneously remarked that the utterances to which our rules had been applied were of a higher storytelling quality than the other utterances, even though we did not explicitly ask them to judge this.

### V. Implementation

The storytelling prosody rules described in Section IV were implemented in a module that creates storytelling speech, based on the output of the Dutch Fluency text-to-speech system. The rules modifying intensity could not be implemented, because Fluency does not allow any control over intensity. Fig. 5 shows an overview of the generation process; below we briefly discuss the steps that are involved.

The input of the module consists of a text with mark-up indicating where storytelling effects should be applied. We extended the Speech Synthesis Markup Language (SSML) [20] with tags specifying the speaking style, i.e., neutral or narrative (= storytelling); indicating the location of sudden and increasing climaxes; and indicating adjectives/adverbs that need to be ‘stretched’. An annotated example text is shown below.

---

2Interestingly, [13] found that liveliness was muted in synthetic speech as compared with the original natural speech. This finding, which stands in contrast to ours, may be specific for their unit selection approach.
VI. EVALUATION

The traditional approach to evaluation of expressive (emotional) speech synthesis is to present subjects with a number of synthesized utterances that are neutral in content, and have them make a forced choice between emotion categories, selecting the category which they think best matches the prosody of each utterance (see [2]–[4] and [6]–[8]). However, it has been argued that perception in this type of experiments is not very accurate, because subjects lack the additional cues that are present in natural situations [21]. Also, for applications that make use of speech synthesis, the question whether the prosody ‘fits in’ with the message is more important than the question whether subjects can categorize the prosody without additional cues [9]. Therefore, following [7], [9]–[11] we decided to evaluate our prosody rules in a realistic context, using utterances that were clearly recognizable as fragments from fairy tales, rather than neutral statements. (We also used this approach in the test determining the best constant values for our prosody rules, described in Section IV-A.)

As material we used 8 short utterances, taken from the stories used in our global analysis. The first five utterances were relatively neutral in content, but typical for fairy tales, e.g., ‘Once upon a time there was a man who was incredibly rich.’ Utterances 6 and 7 contained a sudden climax, and utterance 8 (the only utterance consisting of more than one sentence) contained an increasing climax. A list of all utterances used in the experiment, together with their translations, is given in the Appendix. Two versions were created of each utterance. One version was generated using Fluency, without modifications. The other version was generated using our storytelling speech generation module. SAY HOW MANY SUBJECTS WE USED. THEY WERE NOT EXPERTS ON SPEECH SYNTHESIS. The subjects were divided into two groups, so that each subject judged only one version of each utterance: either the neutral or the storytelling version. The fragments were presented to them in a randomized order.

After a short introduction to the experiment, the subjects listened to a short speech sample, intended to let them get used to the synthetic speech. After that they were presented with the eight synthesized utterances. For each utterance, the subjects were asked to rate its storytelling quality, naturalness and expression of suspense on a 5-point scale. The questions we asked them were, "How do you judge the quality of storytelling of this speaker?" (1 = very bad, 5 = excellent), "How do you judge the naturalness of the fragment?" (1 = very unnatural, 5 = very natural) and "How suspenseful do you perceive the fragment?" (1 = not suspenseful, 5 = very suspenseful). They also had the option to provide free comments about each utterance.

Table I gives the results of this evaluation. Average ratings for the neutral versions are given in the A columns, ratings for the storytelling versions are given in the B columns.

Overall, the versions generated using our storytelling speech
module were judged to have higher storytelling quality than the prosodically neutral versions, with statistical significance of $p < .05$ (using t-test and Mann-Whitney test) only for utterance 1. In addition, the storytelling versions were judged to be more suspenseful than the neutral versions, with statistical significance for utterances 1, 6 and 8. Interestingly, the higher suspense rating held for all utterances, even for utterances 1-5, which were not particularly suspenseful in content, and to which no suspense rules had been applied. So apparently, the general storytelling speaking style already adds some suspense to relatively neutral utterances. The additional suspense rules used in utterances 6-8 do seem to increase the feeling of suspense even further, however, because for these utterances the difference in suspense with the neutral versions was even more marked.

For some utterances, the naturalness of the storytelling versions was judged to be slightly lower than that of the neutral versions. No significant differences were found, however. Probably the lower rating for naturalness is caused by our relatively crude manipulations, which introduced some clearly audible artifacts in the generated speech. In their comments, subjects indicated that they found some of the storytelling effects slightly over-exaggerated; in particular, the stretching of the vowels in certain adjectives. However, in spite of these imperfections, the manipulated fragments mostly received positive comments, with subjects describing the style as ‘graceful’ and ‘interesting’. The regular text-to-speech output was referred to as ‘dull’ and ‘flat’.

VII. General Discussion and Future Work

The prosody rules we use in our storytelling speech generation system are based on the analysis of a fairly small number of speech recordings. To establish how general our findings are, we need to analyze more speech material. Preferably, we should make our own recordings of different speakers telling the same story, so that we can make reliable comparisons. A larger set of speech data might also allow us to reformulate our prosody rules in a probabilistic fashion, which would give rise to more natural variations in the output. In addition, naturalness might be improved by also taking voice quality differences into account. The diphone-based text-to-speech system we have been using so far does not allow this, however. Another factor that may have influenced our results is that there may be minor prosodic differences between the newsreader speech that we based our analysis on and the output of the text-to-speech system we used. To investigate such potential differences, a comparison between the two should be carried out, e.g., by synthesizing the transcripts of some news fragments and comparing the results with the original speech.

The two types of suspense we distinguished in our analysis are tied to the occurrence of climactic events in the story, either expected (increasing climax) or unexpected (sudden climax). In addition to this, other forms of suspense might be distinguished. For instance, [22] distinguish different suspense categories based on the morphological function of the different scenes in a story (e.g., ‘departure of the hero’, ‘struggle with the enemy’). It will be interesting to investigate whether these types of suspense have perceptible prosodic correlates.

The evaluation of our storytelling speech generation module has shown encouraging results. However, we cannot draw any strong conclusions from these, due to the small number of subjects and stimuli used in the evaluation experiment. To get more significant results, we will need to perform a larger-scale evaluation, possibly also including natural speech material as an extra point for comparison of the synthetic speech results (cf. [13]). Also, it would be interesting to perform evaluations with children, as they are the main audience for fairy tales and have been shown to appreciate large manipulations of pitch and duration [10].

The storytelling speech generation module described here has not yet been integrated in our target application, the Virtual Storyteller [1]. The storytelling system should be able to generate good quality output texts before justice can be done to the prosody generated by our storytelling speech generation module. Therefore, we have recently developed an improved version of the language generation component of the Virtual Storyteller [23] and we are currently working on integrating this component into our system. The next step will be to extend this component so that it can automatically generate the tags required by our speech generation module.

For stories that are automatically created by the Virtual Storyteller or other story generation systems [15], [17] the underlying story structure and meaning is already known, and it should be possible to use this information for automatically adding the tags required by the storytelling speech generation module. Another option would be to use our module as a story reading system that takes existing story texts as input. Of course, automatically determining which parts of a plain text are suspenseful or need extra emphasis is still an unsolved issue. Still, even in the absence of specific tags the global storytelling speaking style could be used to synthesize these texts in a more expressive way than a standard text-to-speech system would. The global storytelling style is expected to be relevant also for other applications that require highly expressive speech, such as applications aimed at children [10]. It may be particularly suitable for children with language-related disabilities, because expressive speech provides special benefits for this group (see [21] for arguments).

VIII. Conclusions

In this paper we have described how a storytelling speaking style can be achieved by modifying the prosody of utterances produced by a text-to-speech system. We have evaluated our work in a small perception experiment, where subjects rated the output of our storytelling speech generation module on storytelling quality, naturalness and expression of suspense. The experiment showed that some of our modifications lead to effects that were perceived as over-exaggerated, and in some cases the speech quality slightly degenerates. Still, on the whole subjects preferred the modified speech over the original, neutral version, as it is livelier and more fitting for fairytale stories. Although the number of participants in the evaluation was too low to draw any strong statistical conclusions, we find
these results encouraging. The practical use of our work is not limited to story generation systems like the Virtual Storyteller. It can also be used for synthesizing existing stories, or in various other applications requiring expressive speech.

APPENDIX
1) *Er was eens een man die geweldig rijk was.* (‘Once upon a time there was a man who was incredibly rich.’)
2) *Dan zat hij in een grote stoel met een schitterend geborduurde rug.* (‘Then he sat in a big chair with a beautifully embroidered back.’)
3) *Hij was de rijkste man van het hele land en toch was hij niet blij en gelukkig.* (‘He was the richest man in the entire country and still he wasn’t cheerful and happy.’)
4) *Die baard maakte hem zo afschuwelijk lelijk dat iedereen op de loop ging zodra hij in de buurt kwam.* (‘That beard made him so terribly ugly that everybody ran away when he came near.’)
5) *Als ze maar dachten dat ze ergens muziek hoorden dan bewogen ze zich sierlijk op de maat van die muziek.* (‘Whenever they thought they heard music somewhere they moved gracefully with the rhythm of the music.’)
6) *Hij rende zo hard als hij kon maar toen struikelde hij over zijn eigen benen.* (‘He ran as fast as he could but then he stumbled over his own legs.’)
7) *Hij wilde zich omkeren en toen klonk er plotseling een harde knal.* (‘He wanted to turn around and then suddenly there was a loud bang.’)
8) *Stap voor stap kwam hij dichterbij. Toen hij haar dicht genoeg genaderd was greep hij haar bij haar keel en toen bleek ze plotseling verdwenen.* (‘Step by step he came closer. When he had come close enough to her, he grabbed her throat and then suddenly she disappeared.’)

ACKNOWLEDGMENT
We would like to thank our three anonymous reviewers for their detailed and very useful comments on the first version of this paper.

REFERENCES

Mariët Theune Biography text here.

Koen Meijs Biography text here.

Dirk Heylen Biography text here.

Roeland Ordelsma Biography text here.