Generating Expressive Speech for Storytelling Applications

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

***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. An evaluation of our storytelling speech generation system 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 and 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. Originally, this was done in a crude fashion, using fixed templates for language generation in combination with a standard text-to-speech system. However, we have recently been 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 the creation of a general storytelling speech style, and on the prosodic expression of suspense. Our approach has been to perform an analysis of human storytelling speech and based on this, design a set of rules that modifies the prosodic parameters produced by a Dutch text-to-speech 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.

This paper is structured as follows. In Section II, we give an overview of related work. In Section III, we describe the prosodic patterns we observed in human storytelling speech, focusing on global speech style and the expression of suspense. In Section IV, we discuss our rules for converting neutral speech to storytelling speech, and we describe their implementation in Section V. In Section VI, we describe the evaluation of our storytelling speech generation module. We end with a discussion and conclusion in Section VII and Section 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 nonemotional 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 [10]. For a training application in the military domain, [11] used a limited domain

Manuscript received January 1, 2005; revised January 13, 2006. The work of M. Theune was supported by the Netherlands Organization for Scientific Research (NWO) under Grant 532.001.301. The authors participate in the EU Network of Excellence HUMAINE. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerard Bailly.

M. Theune, D. Heylen, and R. Ordelman are with the Human Media Interaction Group, University of Twente, 7500 AE Enschede, The Netherlands (e-mail: m.theune@utwente.nl; d.k.j.heylen@utwente.nl; r.j.f.ordelman@utwente.nl).

K. Meijs was with the University of Twente, 7500 AE Enschede, The Netherlands. He is now with Arinso Nederland B.V., 3001 DD Rotterdam, The Netherlands (e-mail: koen@nephila.nl).

Digital Object Identifier 10.1109/TASL.2006.876129

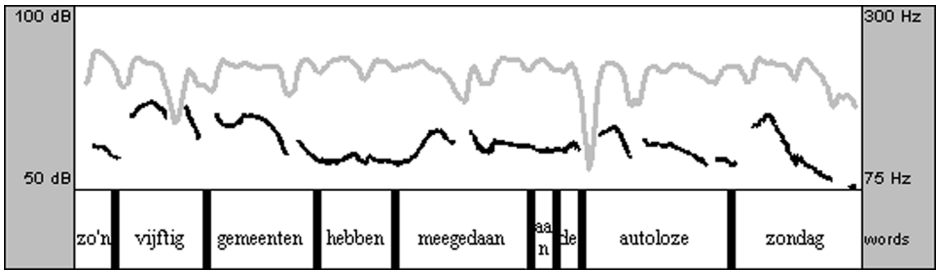

Fig. 1. Newsreader intensity and pitch. (Translation: "Around 50 municipalities have participated in the carless Sunday.")

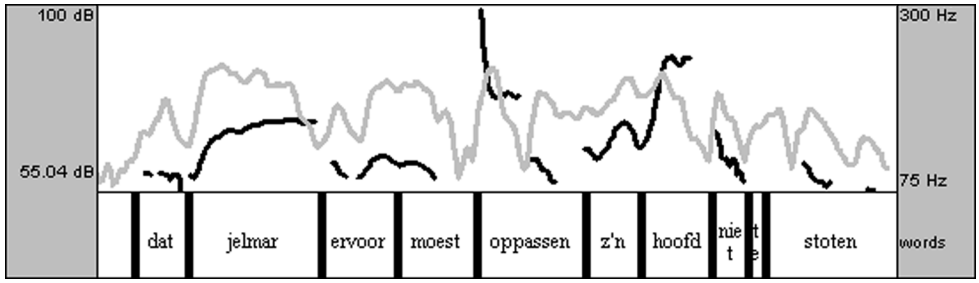

Fig. 2. Children's storyteller intensity and pitch. (Translation: "that Jelmar had to be careful not to bump his head.")

unit selection approach to generate shouted commands, spoken commands, shouted conversation, and spoken conversation. To generate speaking styles for different dialog contexts, [12] trained prosody models for expressing good news, bad news, and questions. As part of a pilot study into expressive speech synthesis, [13] 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 has been 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 investigation, we used existing recordings of children's fairy tales, narrated by professional Dutch voice actors. We randomly selected five stories of five to twelve minutes long, read by three different male storytellers. As comparison material, we used recordings of four short (30 s to 1.5 min) 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] In the following, we describe our general observations of the 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 the storyteller and newsreader speech, we observed that the storytellers used much more variation in pitch and intensity than the newsreaders. This is illustrated in Figs. 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 to take longer pauses, particularly between sentences. Finally, the storytellers sometimes added extra emphasis to certain adjectives and adverbs by increasing their pitch and duration. For example, "A *looong* 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 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.

The first type is the *sudden climax*: an unexpected dramatic moment in the story, such as a startling revelation or a sudden momentous event. Typically, in our material, such climactic events are announced by a steep increase of intensity and pitch on the keyword introducing the climax ("then," "suddenly,"

[1]This approach presupposes that the default speech synthesis is roughly equivalent to human newsreading style. We did not check this assumption.

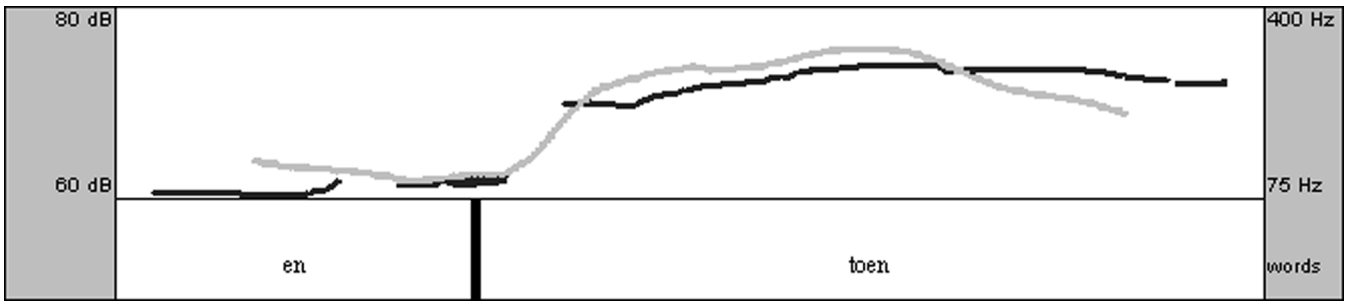

Fig. 3. Sudden climax intensity and pitch. (Translation: "and then.")

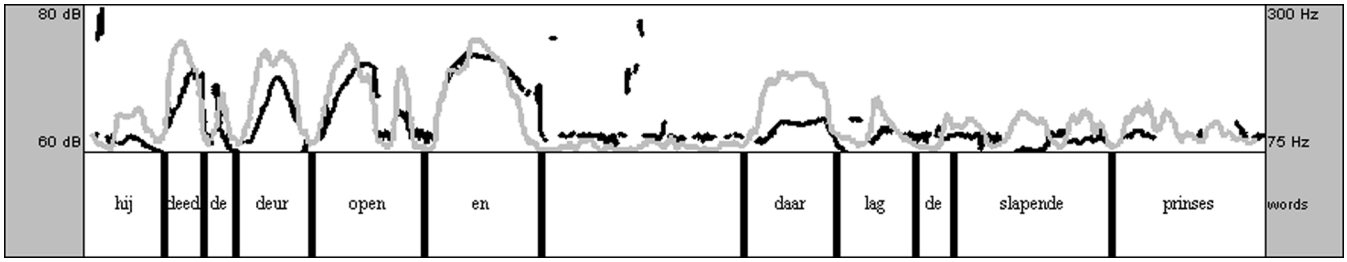

Fig. 4. Increasing climax intensity and pitch. (Translation: "He opened the door and . . . there was the sleeping princess.")

"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 heighten the suspense by a gradual increase in pitch and intensity, accompanied by a decrease in tempo. They typically add 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 the 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 discuss the main rules involved in creating a global storytelling style and the two types of suspense.

A. Method

Our approach in creating the storytelling prosody rules was as follows. First we made a global rule design, based on our observations from the previous section. Then, a few representative samples from our speech material were analyzed in detail, using the speech analysis tool *Praat* [18]. The selected fragments did not contain any "special effects" such as emotional speech or mimicking the voice of a character. We avoided 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 because we would not be able to replicate these using Fluency.

Based on the analysis of the selected speech fragments, we determined the possible range of the 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 speech pairs used in the test were created as follows. First, we synthesized 23 fragments from our storytelling material using Fluency text-to-speech. Of these, five fragments expressed a sudden climax, six expressed an increasing climax, and 12 were neutral in content (i.e., nonsuspenseful). Then, we manipulated the result using *Praat* scripts [18] simulating the prosody rules, with different values for the constants we wanted to test. We made two versions of each fragment, which only differed from each other with respect to the value of one constant. We also included some nonmanipulated versions in the test. The constant values which scored best in this experiment were used in our prosodic rules.

In the following, we discuss the 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

We designed the following rules 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 newsreaders, 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 most commonly used in storytelling. (See [19] for an analysis of storytelling pitch contours.) The rule looks as follows:

$$y'(t) = y(t) \cdot \left(1 + a * \sin\left(\frac{t - t_1}{t_2 - t_1} * 0.5\pi + 0.25\pi\right)\right) \quad (1)$$

where

t $\in [t_1, t_2]$;
y original pitch value y as a function of t;
y' modified pitch value;
a desired maximum pitch increase divided by average pitch.

To determine the desired maximum pitch increase, we analyzed three storytelling fragments (total length 17.4 s). On the syllables carrying a sentence accent, we found pitch increases between 4 and 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. Moreover, fragments with a pitch increase higher than 60 Hz sounded too unnatural to even include in the evaluation. Therefore, we only tested the range between 30 and 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 and 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. We found 2 dB as the best value.

With regard to **tempo**, we analyzed a few storytelling and newsreading speech fragments of about five sentences each, and found average speaking rates of 3.0–3.6 syllables per second (sps) for storytelling, against 5.8 sps for news reading. In our perception test, we tested speaking rates of 3.0 and 3.6 sps. A rate of 3.6 sps was found to be most natural for storytelling, so we used that for the general storytelling speaking style.

In addition to the general slowdown in tempo, we observed a **duration** increase in the accented vowel of certain words (typically, adjectives and adverbs indicating some extreme value of a property). For two such words found in our material, we measured the duration of the accented vowels, and found that they were stretched to 1.4 and 1.8 times their average duration. Our perception test included only one fragment for which vowel stretching was appropriate (similar to the example in Section III-A). For this fragment, the subjects indeed preferred the version with a 50% vowel duration increase over the version without increase in vowel duration.

Finally, in our speech material, we found differences in **pause** length between storytellers and newsreaders that were larger than predicted by the global difference in tempo. In the fragments used to determine average speech tempo, we found an average pause length of 0.4 s at phrase breaks and of 1.3 s between sentences. (For comparison: in the newsreader speech material, we found average pauses of 0.3 and 0.5 s, respectively.) We added a rule fixing the pause lengths at the found values, without testing them.

C. Rules for Sudden Climax

The rules used to express a *sudden climax* apply to the accented syllable in the word announcing the climax (e.g., "then"), represented by the time domain [t_1, t_2]. 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 fragment 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 analyzed 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.

The 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 found in our test was 6 dB.

Finally, as in the rules for global storytelling style for single words indicating extreme values, 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 revelation (e.g., "He opened the door and then—"). In the rules below, this part is indicated by [t_1, t_2]. In the second part, the actual revelation takes place ("—there was the sleeping princess"). This part is indicated by [t_2, t_3].

In [t_1, t_2] we gradually apply a pitch increase to the accented syllables, indicated by the interval [s_i, s_j]. The corresponding rule is shown in (2).

$$y'(t) = y(t) \cdot \left(1 + a_{ic} * \sin\left(\frac{t - s_i}{s_j - s_i} * 0.5\pi + 0.25\pi\right)\right) \quad (2)$$

where

t $\in [s_i, s_j]$;
y original pitch value y as a function of t;
y' modified pitch value;
a_{ic} desired maximum pitch increase in the first part of an increasing climax [$dmpi_{ic}$ in (3)] divided by average pitch.

Because we want to gradually enlarge the pitch increase of each accented syllable [s_i,s_j] 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

$$dmpi_{ic} = p_1 + (p_2 - p_1) * \frac{s_i - t_1}{t_2 - t_1} \quad (3)$$

where

$dmpi_{ic}$ desired maximum pitch increase;
p_1 desired pitch increase at t_1;
p_2 desired pitch increase at t_2.

To determine the constant values to be used for the increasing climax, we only analyzed one storytelling speech fragment in detail: the one shown in Fig. 4. In this fragment, we found a pitch increase of 100 Hz at t_1 and of 130 Hz at t_2. Since these values did not sound acceptable when we reproduced them for our experiment, we decided to shift down the value ranges to 25–50 Hz for p_1 and to 60–80 Hz for p_2. The perception test resulted in 25 Hz as the best value for p_1, and 60 Hz for p_2. Therefore, 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 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, toward a maximal duration increase of 150% at t_2. At t_2 our rules insert a 1.04 s pause, just before the revelation of the climactic event 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 modification. Except for the pitch values used in (3), all constants used in the increasing climax rules were based on the values found in the single analyzed speech fragment, and were not tested in the perception experiment.

E. Discussion

If we look at the outcomes of the perception test, 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.[2] Another possible explanation is that the more extreme modifications, using the maximum values of the constants 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 represent a step in the right direction, as some participants spontaneously remarked that the speech fragments to which our rules had been applied were of a higher storytelling quality than the other fragments, even though we did not explicitly ask them to judge this.

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

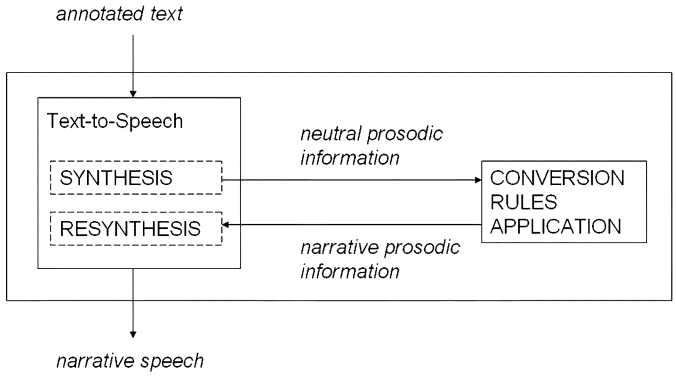

Fig. 5. Storytelling speech generation process.

V. Implementation

Our storytelling prosody rules have been 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. The steps in the storytelling speech generation process are shown in Fig. 5. The input is a text with mark-up indicating where storytelling effects should be applied. We extended the Speech Synthesis Markup Language (SSML) [20] with a tag specifying the speaking style, i.e., neutral or narrative (= storytelling). We also added tags indicating the location of sudden and increasing climaxes, and of adjectives/adverbs in need of "stretching." In terms of the multilevel extension of SSML for expressive speech proposed by [13], our tags are located at the middle level: they carry linguistic information and can be added during language generation. The alternative would have been to use low level numerical tags that directly specify the required prosodic modifications. However, using higher level tags allows us to change our prosodic rules (changing the translation of the tags to prosodic parameters), without any changes in language generation. A marked-up example is shown here.

<speak>
<style type=narrative/>
<s> The beard made him look <extend> so </extend> ugly that everyone ran away when they saw him. </s>
<s> He wanted to turn around <climax type=sudden> and then </climax> there was a loud bang. </s>
<s> Bluebeard raised the big knife, <climax type=increasing> he wanted to strike and <climax_top/> there was a knock on the door. </climax> </s>
</speak>

The input text is sent to the text-to-speech system, which has been set to produce prosodic information instead of an actual speech signal. This information takes the shape of a list of phonemes with their duration and pitch values. This list is sent to the prosody conversion module, which uses the tags in the input text to look up which rules should be applied to which phonemes. The acoustic features of the relevant phonemes are then modified according to the rules. The suspense rules are applied to the neutral prosody as provided by the text-to-speech system; they are not added on top of the effect of the global storytelling style. In other words, the global storytelling style only applies to the "neutral" (nonclimactic) parts of the story. The only exception to this is the tempo, which is set for the whole text, including the climactic parts (there are no suspense rules that influence tempo). The modified prosodic information is returned to the text-to-speech system, which uses it to create a speech signal.

VI. Evaluation

The traditional approach to evaluating 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 such 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, it is more important to know whether the prosody "fits in" with the message than whether subjects can categorize the prosody without additional cues [9]. Therefore, following [9]–[12], we evaluated 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 perception test described in Section IV-A.)

As material, we used eight short transcripts of fragments from our storyteller speech material. The first five fragments were relatively neutral in content, but typical for fairy tales, e.g., "Once upon a time there was a man who was incredibly rich." Fragments 6 and 7 contained a sudden climax, and fragment 8 (the only fragment consisting of more than one sentence) contained an increasing climax. A list of all fragments used in the experiment, together with their translations, is given in the Appendix. Two versions were created of each fragment. One version was generated using Fluency, without modifications. The other version was generated using our storytelling speech generation module. The experiment was performed on line, with 30 subjects who were not experts on speech synthesis. The subjects were divided into two groups, so that each subject judged only one version of each fragment: 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 fragments. For each fragment, the subjects were asked to rate its storytelling quality, naturalness and expression of suspense on a five-point scale. The questions we asked them were, "How do you judge the quality of storytelling of this speaker?" $(1 = \text{very bad}, 5 = \text{excellent})$, "How do you judge the naturalness of the fragment?" $(1 = \text{very unnatural}, 5 = \text{very natural})$ and "How suspenseful do you perceive the fragment?" $(1 = \text{not suspenseful}, 5 = \text{very suspenseful})$. They also had the option to provide free comments about each fragment. Our hypotheses were that the manipulated fragments would score higher on storytelling quality and suspense than the neutral fragments, but lower on naturalness because of our relatively crude manipulations, which introduced some clearly audible artifacts in the generated speech.

TABLE I
Evaluation Results (A = Neutral, B = Storytelling Style)

	Storytelling		Naturalness		Suspense	
Fragment	A	B	A	B	A	B
1	3.0	3.9*	2.8	3.8**	2.1	3.6**
2	2.9	3.1	3.1	2.8	2.3	2.6
3	2.9	3.3	2.6	2.7	2.2	2.7
4	2.9	3.7*	2.7	3.3	1.9	2.9**
5	2.6	2.9	2.6	2.0*	1.9	1.9
6	3.0	3.3	2.4	3.0	2.0	3.1**
7	3.1	3.2	3.0	3.2	2.5	3.0
8	2.9	2.8	3.0	2.8	2.2	3.7**
Total average	2.9	3.2**	2.7	2.9	2.1	2.9**

* $p < .05$, ** $p < .01$ for one-tailed Mann-Whitney test

Table I gives the results of the 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 scored higher on storytelling quality than the prosodically neutral versions, with statistical significance of $p < .01$ (using a one-tailed Mann–Whitney test). In addition, the storytelling versions were judged to be more suspenseful than the neutral versions, also with a significance of $p < .01$. Interestingly, the higher suspense scores for the manipulated versions held not only for fragments 6–8 (to which our suspense rules had been applied), but also for fragments 1–4, even though these fragments were not particularly suspenseful in content, and the suspense rules had not been applied to them. Apparently, the general storytelling speaking style by itself already adds some suspense to relatively neutral fragments.

As expected, for some fragments the naturalness of the storytelling versions was judged to be lower than that of the neutral versions. For other fragments, however, the opposite was the case. From the subjects' free comments, we get the impression that low ratings for naturalness had more to do with misplaced sentence accents than with our manipulations. Overall, we found no significant difference between the manipulated and the nonmanipulated versions for naturalness.

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 prosodic rules we use in our storytelling speech generation system are based on the analysis of a 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 taking voice quality differences into account. The diphone-based text-to-speech system we have been using so far does not allow this, however. Neither does it allow us to use any rules modifying intensity. Another factor that may have influenced our results is that there may be minor prosodic differences between the newsreader speech 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. In the future, we would like to perform a larger scale evaluation, possibly also including natural speech material as an extra point for comparison (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]. Before justice can be done to the prosody generated by our storytelling speech generation module, the storytelling system should be able to generate good quality output texts. Therefore, we have recently developed an improved version of the language generation component of the Virtual Storyteller (see [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 such as [15] and [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. Automatically determining which parts of a plain text are suspenseful or need extra emphasis is still an unsolved issue, but 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. Conclusion

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 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 are perceived as over exaggerated, and in some cases, the speech quality slightly degenerates. Still, on the whole our subjects preferred the modified speech over the original neutral version, as it is livelier and more fitting for fairytale stories. Our evaluation experiment was performed on a small scale, but we find its 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 everyone 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

The authors would like to thank the anonymous reviewers for their detailed and very useful comments on the first version of this paper and E. Herder for his help with the statistics.

REFERENCES

[1] M. Theune, S. Rensen, R. op den Akker, D. Heylen, and A. Nijholt, "Emotional characters for automatic plot creation," in *Proc. TIDSE 2004: Technologies for Interactive Digital Storytelling and Entertainment*, Darmstadt, Germany, Jun. 2004, pp. 95–100. Springer LNCS 3105, 2004.

[2] J. Cahn, "The generation of affect in synthesized speech," *J. Amer. Voice I/O Soc.*, vol. 8, pp. 1–19, 1990.

[3] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Commun. 16*, pp. 369–390, 1995.

[4] M. Bulut, S. Narayanan, and A. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," in *Proc. 7th Int. Conf. Spoken Language Process. (ICSLP 2002)*, Denver, CO, Sep. 2002, pp. 1265–1268.

[5] A. Black, "Unit selection and emotional speech," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1649–1652.

[6] S. Mozziconacci, "Speech variability and emotion: production and perception," Ph.D. dissertation, Univ. Eindhoven, Eindhoven, The Netherlands, 1998.

[7] J. M. Montero, J. Gutiérrez-Arriola, S. Palazuelos, E. Enríquez, S. Aguilera, and J. M. Pardo, "Emotional speech synthesis: From speech database to TTS," in *Proc. 5th Int. Conf. Spoken Language Process. (ICSLP)*, vol. 3, Sydney, NSW, Australia, Nov./Dec. 1998, pp. 923–926.

[8] I. Iriondo, F. Alías, J. Melenchón, and M. Angeles Lorca, "Modeling and synthesizing emotional speech for Catalan text-to-speech synthesis," in *Proc. Workshop Affective Dialogue Syst.*, Kloster Irsee, Germany, Jun. 2004, pp. 197–208. Springer LNAI 3068, 2004.

[9] M. Schröder, "Dimensional emotion representation as a basis for speech synthesis with nonextreme emotions," in *Proc. Workshop Affective Dialogue Syst.*, Kloster Irsee, Germany, Jun. 2004, pp. 209–220. Springer LNAI 3068, 2004.

[10] D. House, L. Bell, K. Gustafson, and L. Johansson, "Child-directed speech synthesis: Evaluation of prosodic variation for an educational computer program," in *Proc. Eurospeech*, Budapest, Hungary, Sep. 1999, pp. 1843–1846.

[11] W. L. Johnson, S. Narayanan, R. Whitney, R. Das, M. Bulut, and C. LaBore, "Limited domain synthesis of expressive military speech for animated characters," in *Proc. IEEE Speech Synthesis Workshop*, Santa Monica, CA, Sep. 2002.

[12] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to ⟨ahem⟩ expressive speech synthesis," in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, Jun. 2004, pp. 79–84.

[13] E. Eide, R. Bakis, W. Hamza, and J. Pitrelli, "Toward Synthetic Expressive Speech," in *Text to Speech Synthesis: New Paradigms and Advances*, S. Narayanan and A. Alwan, Eds. Upper Saddle River, NJ: Prentice-Hall, 2004Text to Speech Synthesis: New Paradigms and Advances, pp. 219–250.

[14] J. Fackrell, H. Vereecken, J. Buhmann, J.-P. Martens, and B. Van Coile, "Prosodic variation with text type," in *Proc. 6th Int. Conf. Spoken Language Process. (ICSLP)*, vol. 3, Beijing, China, Oct. 2000, pp. 231–234.

[15] A. Silva, G. Raimundo, C. de Melo, and A. Paiva, "To tell or not to tell. . . Building an interactive virtual storyteller," in *Proc. AISB Symp. Language, Speech and Gesture for Expressive Characters*, Leeds, U.K., Mar./Apr. 2004. .

[16] A. Silva, M. Vala, and A. Paiva, "The Storyteller: Building a synthetic character that tells stories," in *Proc. Workshop Multimodal Communication and Context in Embodied Agents*. Montréal, QC, Canada, May/Jun. 2001, pp. 53–58.

[17] N. Braun, T. Rieger, and M. Dietz, "VR-NaSty: VR character narrator with story-based suspense support," in *Proc. 12th Int. Conf. Central Europe Comput. Graphics, Visualization, Comput. Vision (WSCG)*, Plzen, Czech Republic, Feb. 2004, pp. 29–32.

[18] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (v. 4.3.04)," Univ. Amsterdam, Amsterdam, The Netherlands, http://www.praat.org/.

[19] E. Klabbers and J. van Santen, "Clustering of foot-based pitch contours in expressive speech," in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, Jun. 2004, pp. 73–78.

[20] Speech synthesis markup language version 1.0. W3C recommendation, Sept. 2004.

[21] J. van Santen, L. Black, G. Cohen, A. Kain, E. Klabbers, T. Mishra, J. de Villiers, and X. Niu, "Applications of computer generated expressive speech for communication disorders," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1657–1660.

[22] N. Braun and T. Rieger, "Expressiveness generation for virtual characters, based on the principles of suspense progression and narrative conflict," in *Proc. Int. Workshop Mobile Computing (IMC)*, Rostock, Germany, Jun. 2003, pp. 113–120.

[23] F. Hielkema, M. Theune, and P. Hendriks, "Generating ellipsis using discourse structures," in *Proc. ESSLLI Workshop Cross-Modular Approaches to Ellipsis*, Edinburgh, U.K., Aug. 2005, pp. 37–44.

Mariët Theune studied computational linguistics at the University of Utrecht, Utrecht, The Netherlands, and received the Ph.D. degree from the University of Eindhoven, Eindhoven, The Netherlands, in 2000, where she worked on concept-to-speech generation.

Since 2001, she has been an Assistant Professor at the University of Twente, Enschede, The Netherlands, on various research projects, including automatic story generation. Her main research involves information presentation by embodied conversational agents (ANGELICA project). She is also involved in the IMOGEN project, focusing on answer presentation in a QA system.

Koen Meijs studied computer science at the University of Twente, Enschede, The Netherlands, and received the M.Sc. degree from the University of Twente in 2004, specializing in language and speech technology. His thesis subject was the generation of narrative speech.

During his years at the university, he started a company specialized in building financial web applications. Since 2004, he has continued working on this at Arinso Nederland B.V., Rotterdam, The Netherlands, a company building solutions in the area of human resource management.

Dirk Heylen studied German philology, computer science, and computational linguistics at the University of Antwerp, Antwerp, Belgium, and received the Ph.D. degree from the University of Utrecht, Utrecht, The Netherlands. His dissertation was on categorial grammar.

Since 1998, he has been an Assistant Professor at the University of Twente, Enschede, The Netherlands. His main subject is affective interactions, between both humans and humans and machines. Currently, he is working on emotion and its verbal and nonverbal manifestations in the IST-Project AMI, and as part of the HUMAINE network.

Roeland Ordelman studied psycholinguistics at the University of Utrecht, Utrecht, The Netherlands, and received the Ph.D. degree from the University of Twente, Enschede, The Netherlands, in 2003, where he worked on Dutch speech recognition for spoken document retrieval.

Since 2003, he has been a Postdoctoral Researcher at the University of Twente. He works on large-vocabulary speech recognition in the meeting domain and on the annotation and detection of emotions in meeting data in the AMI Project. He is also involved in the Dutch MultimediaN program, focusing on context-dependent modeling for Dutch large-vocabulary speech recognition.