Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis

Raúl Montaño[1], *Francesc Alías*[1], *Josep Ferrer*

[1] Grup de Recerca en Tecnologies Mèdia, La Salle, Universitat Ramon Llull, Barcelona
{raulma;falias;st15228}@salle.url.edu

Abstract

The generation of synthetic speech with a certain degree of expressiveness has been successful for some particular applications or speaking styles (e.g. emotions). In this context, there is a particular speaking style with subtle speech nuances that may be of great interest for delivering expressive speech: the storytelling style. The purpose of this paper is to define a first step towards developing a storytelling Text-to-Speech (TTS) synthesis system by means of modelling the specific prosodic patterns (pitch, intensity and tempo) of this speaking style. We base our analysis of a tale in Spanish on discourse modes present in storytelling: narrative, descriptive and dialogue. Moreover, we introduce narrative situations (neutral narrative, post-character, suspense and affective situations) within the narrative mode, which are analysed at the sentence level. After grouping the sentences into modes and narrative situations, we analyse their corresponding prosodic patterns both objectively (via statistical tests) and subjectively (via perceptual test considering resynthesized sentences). The results show that the statistically validated prosodic rules perform equally (or even better) than the original prosody in most sentences.

Index Terms: storytelling, prosodic analysis, narrative situations, TTS, Harmonic plus Noise Model

1. Introduction

Storytelling speaking style has been studied following quite different approaches for the analysis of the specific characteristics of stories and tales. For example, in [1] the authors analysed storytelling according to a common structure of tales (title, exposition, triggering event, a series of scenes, ending and epilogue), whereas in [2] the authors divided the tale into three discourse modes [3] (narrative, descriptive and dialogue), concluding that the storyteller prosody is influenced by discourse modes. In [4], tales and news reading styles were prosodically analysed and compared showing prosodic differences between both styles. In contrast to these global approaches, other works addressed the tale analysis at the sentence level. Specific narrative passages (global storytelling speaking style, increasing suspense and sudden suspense) were studied, modelled and synthesized in [5]. On the other hand, some works modelled the whole story following an emotional approach, for both analysis [6] [7] [8] and synthesis purposes [9] [10], while others only considered emotions for the characters of the story [11].

Nevertheless, none of these works offers a complete solution to deal with the prosodic analysis and modelling of the storytelling speaking style at the sentence level oriented to speech synthesis of all the expressive registers of a storyteller. The storyteller is the person narrating the tale, e.g., the story and the situations that the characters are experiencing. Optionally, he/she can interpret all/part of the characters turns too. To that effect, storytellers make use of a wide range of speech variability in order to convey the necessary expressiveness to capture the audience's (generally, children) attention. For example, they may use rhythm changes or include pauses of different duration, add suspense to the voice, use much more variation of pitch and intensity than other speaking styles such as the newsreader speaking style, stretch some words, etc. [5]. Dialogues are also present in many novels and tales because it is a factor that can engage the audience in the story in a greater way as they can read/listen to the characters directly. Moreover, in oral communication the narrator may give different voices and emotional content to different characters to enhance realism and entertainment. In contrast, in the narrative and descriptive modes, more subtle nuances appear to convey the storytelling style.

In this work, we propose a first approach to cope with that issue performing a prosodic analysis of a story narrated by a Spanish storyteller based on storytelling discourse modes [2] [3]. However, we consider that the prosodic analysis should be conducted at the sentence level to capture all the potential expressive registers of a storyteller following a bottom-up approach. To that effect, we introduce new sub-modes (narrative situations) inside the narrative mode to cope with the sentence level analysis, which will be the basis for further synthesis purposes. Finally, we have chosen the main character of the story to analyse the dialogue mode using an annotation scheme based on basic emotions. After a two-phase analysis of the story at hand, the narrative situations are both objectively and subjectively validated by means of statistical significance analysis and a subsequent preliminary perceptual test considering the corresponding synthesis from the extracted prosodic rules. For the dialogue mode, we compare our results with other studies that have analysed basic emotions to observe if emotions in storytelling show equal or specific prosodic patterns. The obtained emotional rules are also tested in the synthesis phase.

This paper is structured as follows. Section 2 reviews related work on the analysis and synthesis of storytelling speaking style. In Section 3, the proposed approach for storytelling speech analysis is described. Next, the prosodic analysis is detailed in Section 4. Then, the perceptual evaluation with synthesis using the extracted prosodic rules is described in Section 5. Finally, some conclusions and future work are present in Section 6.

2. Related Work

The particular challenges of generating storytelling speaking style were discussed in [9], where the authors stumbled upon this problem as the Text-To-Speech (TTS) system of their embodied digital storyteller did not offered the desired expressive-

ness. According to the authors, the lack of flexibility of the considered TTS system was the main problem. Probably, the fact that the prosodic model was based on emotion profiles borrowed from the literature was also a relevant factor, since they are not entirely well-suited for recreating the storytelling speaking style (e.g., different approaches like [5] seemed to obtain better synthetic results). A later work by some of the authors (centred on interactive storytelling) also remarked that a main obstacle in their work was the synthetic quality of their TTS system [12]. In [13], similar conclusions were obtained on a project devoted to give a robot the ability to tell tales to children. The authors claimed that in storytelling there are particular expressive turns, such as different degrees of emphasis, changes of registers and tempo, different characters, etc., that must be included in the synthetic discourse.

As in [9], later works have linked basic emotions with storytelling. Emotional tags were used to analyze a storytelling speech corpus in [6], which led to a certain degree of correspondence with previously reported emotional acoustic profiles in the literature. Nevertheless, some particular contradictory results (as pitch decrease for anger) were also obtained. Moreover, emotional acoustic models borrowed from the literature were only used for characters from stories in [11]. Although the model was preliminary and needed further work, the synthetic results showed that the changed emotional fragments compared to the neutral fragments were mostly noticeably different, and five emotions were accepted at a reasonable rate.

In [5], the authors only modelled global storytelling speaking style and suspense situations (increasing suspense and sudden suspense). The resynthesized speech generated according to the obtained set of prosodic rules obtained good synthetic quality. However, the rules where highly preliminary because of the very small amount of data considered for the analysis (2 sentences for the sudden suspense and 1 sentence for the increasing suspense). Although the authors proposed a 'global storytelling style', we consider that there is still room for further research towards defining a truly general storytelling style.

A high level annotation scheme according to a common structure of tales (title, exposition, triggering event, a series of scenes, refrain and epilogue [14]), was used in [1] to analyse the prosody of the aforementioned tale sections. However, the authors pointed out that an annotation of affect and emotional tags at the sentence level would be necessary to refine their results. Furthermore, a high level prosodic analysis of a tale was also carried out in [2] in order to perform automatic classification of sentences. The authors labelled the text of a tale among narrative mode, descriptive mode and dialogue mode. The authors argued that prosody is used to mark discourse modes.

Taking these works into account, we base our analysis on storytelling discourse modes but going into the sentence level considering our final synthesis goal. Therefore, new sub-modes, denoted as narrative situations, have been defined as explained in Section 3. Then, a series of prosodic rules are extracted and validated (see Section 4).

3. Narrative situations and character emotions

How should one deal with the classification of text and expressive content in storytelling? Trying to categorize each sentence of a story into one specific basic or secondary emotion or attitude does not seem to be a very good idea. First, relating the narrative style to emotions seems inappropriate, as the narrator is not self-experiencing the emotions and it is not his/her intention to simulate them but to engage the audience in the story. Secondly, gathering a representative corpus for each emotional or attitude to look for speech correlates would be thoroughly intractable [15].

The annotation framework that is used in this work for the further generation of synthetic storytelling speech is based on discourse modes [2] [3]. Specifically, among all discourse modes (narrative, descriptive, argumentative, explanatory and dialogue), the fiction literature (storytelling) typically contains the narrative mode, the descriptive mode and the dialogue mode [2] [3]. In the literary field, the narrative mode is mainly used to inform the listener/reader about the actions that are taking place and affect the characters of the story. Therefore, this mode includes a great amount of text that a storyteller (an expressive one at least) conveys in different expressive registers typically at the sentence level.

The story analysed in this work is "Harry Potter and the Philosopher's Stone" read by a Spanish male storyteller[1]. For the indirect discourse, we have analysed the first chapter of the story, whereas for the study of the dialogue mode, we consider the interventions of the main character of the story (Harry Potter) extracted from the whole story. We followed a two-phase analysis method: a linguistic analysis and a subsequent perceptual refinement. The annotation at the sentence level of the text of the story was entrusted to two experts on text classification. They were instructed to classify sentences as descriptive mode, dialogue mode (Harry's interventions) and narrative mode (see phase 1 of Figure 1).

For the narrative mode they were instructed to classify the sentences according to valence sub-modes (neutral, positive and negative sentences), since it is a useful representation for affective situations where the emotional state is not fully defined [16]. They also were asked to classify what we call post-character sentences. These situations correspond to sentences of indirect discourse immediately following a direct discourse (character intervention) with usually a declarative verb on the third person [17]. Sentences where the annotators did not agree (9.1% of the total number of sentences) were discarded for the second phase.

Once the sentences were classified from text, they were presented to two experts on speech technologies to further analysis (see phase 2 of Figure 1). A briefing with some examples of the different categories they had to listen was given beforehand. Since we are interested in modelling the prosody of the narrator, we consider that we cannot limit our annotation scheme to text and structure characteristics of stories and tales. Their observations were that the affective situations (sentences with positive and negative valence) were too heterogeneous perceptually, and needed a refinement based on activation. Therefore, they classified the affective sentences into Positive/Active, Positive/Passive, Negative/Active and Negative/Passive situations. In addition, they noticed that several neutral and negative sentences possessed a certain degree of suspense. These sentences seemed to possess a greater suspense and tension (expressed with a softer voice) typically caused by a strange event or something the characters of the story are unaware of, leaving the audience to think that something important may happen soon. After considering this fact, we decided to take them into account in the acoustic analysis as a new category: suspense situations. With respect to phase 1, 79% of the suspense sentences came

[1] http://www.ivoox.com/podcast-harry-potter-piedra-filosofal-j-k-rowling_sq_f137546_1.html

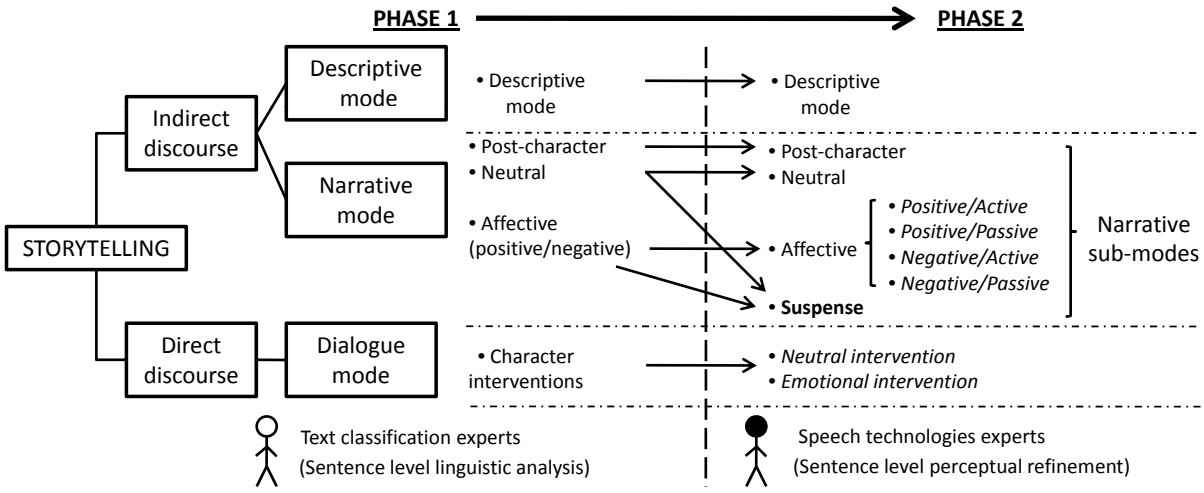

Figure 1: *Diagram representing the followed approach in the present work to analyse the storytelling corpus. Categories in phase 2 are in italics if they were refined from previous categories, in bold if they resulted in new categories and in standard style if they were the same from phase 1.*

Table 1: *Total amount of identified sentences in the speech corpus. Time is expressed as mm:ss.*

Category	# sentences	time
Neutral narrative	46	03:15
Negative/Passive	36	02:34
Negative/Active	30	02:20
Positive/Passive	30	02:22
Positive/Active	31	02:56
Post-character	30	01:03
Suspense	27	01:54
Descriptive mode	30	03:11
TOTAL	260	19:35

(a) Narrative & Descriptive modes

Category	# sentences	time
Neutral	14	00:33
Hot anger	18	00:35
Cold anger	15	00:30
Joy	14	00:15
Sadness	15	00:36
Surprise	12	00:19
Fear	18	00:56
TOTAL	106	03:44

(b) Dialogue mode

from the neutral sentences whereas 21% came from negative sentences (see transition from phase 1 to phase 2 in Figure1). In the related literature, two types of suspense situations (sudden and increasing suspense) present in storytelling have already been modelled [5]. Although some sentences could be related to those kinds of suspense, it was such a small speech corpus that led us to omit these subdivision for the following prosodic analysis.

For the classification of the dialogue mode we use a basic emotions annotation scheme. If a narrator interprets the characters, he/she typically modifies his/her voice into a more exaggerated register of expressions, where full-blown emotions may be manifested [11]. From the gathered corpus, both experts on speech technologies were asked to classify the sentences into six basic emotions (hot anger, cold anger, joy, sadness, surprise and fear) besides a neutral category.

The final collected corpus for the prosodic analysis is detailed in Table 1. We are aware that this is not a very extensive corpus, but it serves as a preliminary step to observe the viability of the approach.

4. Prosodic analysis

In this section, we present the results of the prosodic analysis performed on the sentences collected and labelled after the two-phase process described in Figure 1. We analyse pitch, intensity and tempo. The parameter for modelling tempo is speaking rate (SR) in syllables per second without pauses. Pitch is represented with mean pitch (MP) and pitch standard deviation (PSD) in Hertz while for intensity, we extracted mean intensity (MI) in decibels. We used the speech analysis software Praat [18] to extract these prosodic parameters. In order to maximize the measurement precision of pitch, each sentence received optimal pitch floor and ceiling values computed with the MOMEL plug-in for Praat [19] (manually corrected if needed). On the other hand, the SR was measured with the ADoTeVA Praat plug-in[2]. The segmentation of the speech corpus into words, syllables and phonemes was carried out with the EasyAlign tool [20], and was manually corrected afterwards.

First, we analyze the indirect discourse, i.e., the narrative mode with its associated narrative situations and the descriptive mode. We select the neutral narrative situation as the reference category, so the results of the rest of categories are referenced to this one in terms of relative percentage difference. An independent samples t-test has been performed using the SPSS software to check significant differences with respect to the neutral narrative situation and between all categories with pairwise comparisons. From this statistical analysis we want to determine to what extent categories are different and which parameters are more significant. The p-value used includes a correction when equal variances can not be assumed.

The dialogue mode analysis shows the prosodic results of the basic emotions present in the main character of the story (Harry Potter). We compare these results with other studies that have analyzed acted basic emotions in order to observe if storytelling emotions show equal or specific prosodic patterns.

Table 2: *Averaged results of the indirect discourse mode analysis. Statistical significance tests: * stands for p<0.05, while ** represents p<0.01. No * means no significant difference.*

Narrative mode	MP [Hz]	PSD [Hz]	SR [syll/sec]	MI [dB]
Neutral narrative	104.0	31.0	7.7	71.0
Post-character	-22.3%**	-50.2%**	-12.1%**	-4.3%**
Suspense	-5.7%**	-19.0%**	-10.9%**	-4.3%**
Negative/Passive	-12.9%**	-35.3%**	-8.0%**	-3.8%**
Negative/Active	+7.9%*	-3.3%	-2.2%	+1.2%
Positive/Passive	-8.9%**	-20.2%**	-8.4%**	-2.6%**
Positive/Active	+18.3%**	+24.3%**	-5.5%**	+2.5%**
Descriptive mode	+1.0%**	+10.0%**	-8.7%**	+0.3%

4.1. Indirect discourse results

Prosodic results for the narrative and descriptive modes are shown in Table 2, where they are referenced with respect to the neutral narrative situation. The statistical analysis is also depicted in Table 2 and the rest of statistical significance comparisons are shown in Table 3. The first global conclusion that arises from the results is that, in general, SR is not significantly different across categories. The analysed storyteller shows a fast SR for neutral narrative and Negative/Active situations, whereas in the rest of situations a slow SR is manifested, probably to allow the audience (children, in general) to follow the story.

Post-character sentences obtained the lowest averaged pitch and intensity values, which is in agreement with the perception of the speech technologies experts that, in general, these sentences sounded muffled. The SR has a mid-low value. The suspense situation shows low prosodic parameters too. As stated in [1], it seems necessary at least a low mean intensity to generate intimacy or suspense. These two situations only show significant differences in terms of pitch as it can be observed in row six of Table 3.

The results for the affective situations show that prosodic parameters from active sentences are significantly higher than the parameters from passive sentences with the exception of SR, which do not always follow this behaviour (see rows 16 to 19 in Table 3). These results quite agree with the established consensus in the literature that active sentences entail higher frequency, intensity and speaking rate [16]. Results in Table 2 show that sentences with positive evaluation have slightly higher prosodic values compared to sentences with different evaluation but with the same activation, with the exception of SR, which is lower. Although there are no clear acoustic correlations with valence in the literature, in [16], a higher mean frequency for a male voice was reported for positive valence, just as the results observed in the affective situations results of Table 2. It is worth pointing out that the passive categories have lower MP and MI when compared to the neutral narrative style, while for the active categories the opposite happens. On the other hand, SR for all the affective situations is slower than the neutral narrative situation SR. Finally, the PSD of Positive/Active sentences is the only one that surpasses the neutral narrative category.

Descriptive mode sentences have a higher MP and PSD than the neutral narrative situation whereas mean intensity is not significantly higher. The SR, however, is lower. All this information can be linked to what was perceived while listening to the speech corpus, as the narrator emphasizes certain adjectives and adverbs, which yields to greater pitch variability, and he stretches these words too in order to emphasize. The descriptive mode and the Positive/Active situation are the only categories that show no significant differences in their prosodic patterns (see row 14 in Table 3). One possible explanation is that the narrator when is describing tends to show a cheerful mood, but further investigation may disambiguate both categories.

[2] http://celinedelooze.com/MyHomePage/Praat.html

Table 3: *Results for the independent samples t-test analysis of indirect discourse categories: * stands for p<0.05, while ** represents p<0.01. P-C: Post-Character, PA: Positive/Active, PP: Positive/Passive, NA: Negative/Active, NP: Negative/Passive, SUS: Suspense, DM: Descriptive Mode.*

Compared categories	Test results			
	MP	PSD	SR	MI
P-C vs. DM	**	**	0.26	**
P-C vs. NA	**	**	**	**
P-C vs. PA	**	**	*	**
P-C vs. PP	**	**	0.27	0.13
P-C vs. NP	**	**	0.19	0.75
P-C vs. SUS	**	**	0.70	0.70
SUS vs. PA	**	**	*	**
SUS vs. NA	**	*	**	**
SUS vs. DM	**	**	0.41	**
SUS vs. PP	0.45	0.83	0.41	**
SUS vs. NP	**	**	0.30	0.34
PP vs. NP	**	**	0.99	0.06
PA vs. NA	**	**	0.18	**
PA vs. DM	0.55	0.97	0.20	0.20
NA vs. DM	**	**	**	**
PP vs. PA	**	**	0.33	**
PP vs. NA	**	*	*	**
NP vs. PA	**	**	0.35	**
NP vs. NA	**	**	*	**
DM vs. PP	**	**	0.91	**
DM vs. NP	**	**	0.79	**

Table 4: *Averaged results of the character emotions analysis.*

Emotion	MP [Hz]	PSD [Hz]	SR [syll/sec]	MI [dB]
Neutral	108.0	25.8	7.2	70.0
Hot anger	+82.8%	+112.3%	-20.6%	+9.1%
Cold anger	+42.4%	+69.0%	-16.7%	+4.0%
Joy	+28.9%	+67.6%	-11.2%	+7.2%
Sadness	-11.5%	-28.5%	-21.6%	-3.3%
Surprise	+45.2%	+92.7%	-15.9%	+0.7%
Fear	+29.1%	+27.2%	-2.2%	+5.3%

4.2. Direct discourse results

To analyse part of the dialogue mode present in storytelling we selected the main character (Harry Potter). The narrator interprets Harry without changing his voice too much, but it is noticeable he tries to imitate the voice of a pre-teenager.

Regarding emotional prosodic results (see Table 4), it is remarkable that all the emotions have a slower SR than the character neutral voice. In general, anger, joy, surprise and fear tend to have a faster SR in the literature [21] [22] [23] [24]. However, the difference between joy and happiness is not so clear in the literature. For example, In [22] the authors clearly

separated them and proposed a decrease in tempo for happiness (as in [24]) and an increase for joy. Results of SR from Harry's emotions are the ones which have more conflict when compared to the general literature focused on emotion analysis. As a preliminary conclusion, it seems that storytellers speak slower even in the character emotions (besides the indirect discourse). This can be due to the fact that they need to draw the audience attention and allow them to be able to follow all the delivered information. Thus, in this parameter may be the main difference with respect to more natural or spontaneous emotions.

Hot anger has the most exaggerated values of MP, PSD and MI of all the emotional catalogue. The raise of the mentioned prosodic parameters is quite coherent with previous studies focused on basic emotions [21] [22] [23]. Cold anger has the same changes as hot anger but not so wide. Joy shows the highest MI right after hot anger, and its pitch related values are quite high in general. Sadness is the emotion which has more relationship with the acoustic profiles reported in the literature, as it entails a decrease in all the prosodic parameters [21] [22] [23] [25]. Surprise, which is usually related to an increase of the prosodic parameters with respect to a neutral register, has also relationship with other studies (except for speaking rate as well) [23]. Fear has a relative coherency with the literature. In general, pitch, intensity and speaking rate also increase in fear when compared to a neutral register [22] [23]. From Table 4, we can see that MP, PSD and MI increase. The SR obtained for fear is the highest of all the emotions, almost the same as the one for Harry's neutral voice.

5. Speech synthesis evaluation

The main objective of the synthesis evaluation stage performed in this work is to subjectively validate the rules obtained for the different discourse modes (see Tables 2 and 4), as a complement of the objective statistical analysis performed in Section 4. We evaluate how the extracted prosodic rules perform against the original prosody of the sentences.

We applied these prosodic rules to a randomly selected set of sentences from the corpus at hand using a synthetic female voice obtained with the TTS synthesizer of La Salle R&D. We resynthesized 52 sentences (4 sentences for each category) with the obtained prosodic rules (PR) and the same 52 sentences applying the original prosody (OP) of each sentence. The modifications and final signal resynthesis were done using a MATLAB implementation of Harmonic plus Noise Models (HNM) [26]. In contrast to other implementations where the maximum voiced frequency is allowed to vary [27], the implementation used in this paper is fixed at 5Khz based on [28].

The synthetic results were evaluated using the online TRUE platform [29]. The subjective test was performed by 15 people, from which 9 are male and 6 female with a mean age of 34 (only 5 people are familiar with the field of speech technologies). The perceptual test is designed considering a 5-level CMOS scheme (OP much better, OP better, no difference, PR better and PR much better), and it is composed of 52 comparisons of the same sentence resynthesized with PR and OP, which are compared including the original sentence of the audio book as a reference.

Figure 2 shows the results obtained for the sentences belonging to the indirect discourse. As a general result, it can be observed that the most extreme cases of the 5-level CMOS range are the least chosen options, showing that both transformations (PR and OP) are perceived similarly. However, post-character sentences tend to be preferred when the original prosody is applied. This can be due to the fact that sometimes

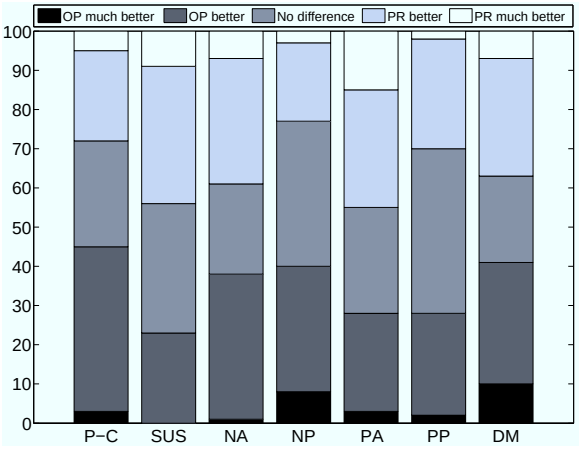

Figure 2: *Percentages bars of the results from the indirect discourse synthesis evaluation.*

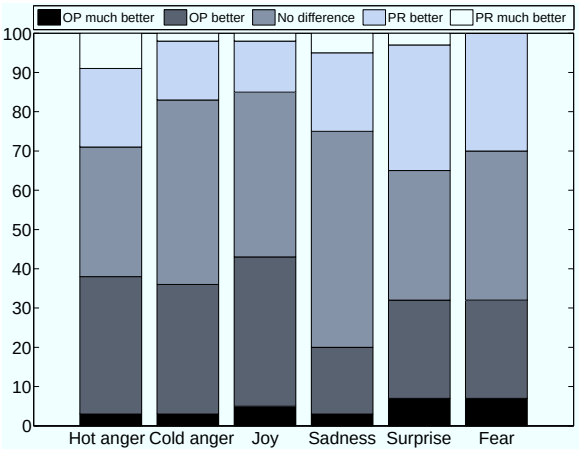

Figure 3: *Percentages bars of the results from the direct discourse synthesis evaluation.*

the narrator maintains emotional traces from the previous character intervention whereas in other post-character sentences he barely is expressive. Suspense sentences have obtained very good results when synthesized with the PR.

The results extracted from the analysis of the emotional sentences of the main character show that most evaluators did not notice clear differences between both prosodic options (see Figure 3), which is a positive result in terms of extracting preliminary general prosodic patterns. However, the PR are not as clearly preferred to the OP (in hot anger, cold anger and joy above all) as in the indirect discourse. This is due to the fact that the emotions show greater prosodic variability and the emotional speech corpus gathered could not be extended.

6. Conclusions

In this paper, we have presented a first approach to cope with the prosodic analysis and modelling of the subtle expressive registers present in the storytelling speaking style at the sentence level. After a linguistic and perceptual analysis of a story based on storytelling discourse modes (narrative, descriptive and dialogue) we have introduced some narrative situations. Next, we have performed a prosodic analysis and extracted preliminary

prosodic rules that have been implemented in a HNM synthesis phase. The outcome of the statistical and synthesis evaluation stages show a first confirmation that there are expressive categories inside the storytelling speaking style that show specific prosodic cues and can be modelled for synthesis purposes. This confirms and extends the conclusions in [5], where specific suspense situations were modelled giving room for further investigation of storytelling expressive registers.

These results encourage us to follow this approach in further studies to look for a *truly* generalizable storytelling speaking style prosodic model. We plan to include more narrators or the same narrator telling a similar story, cross-language analysis, or other forms of stories such as short fairy tales. Short fairy tales tend to have a more common structure than novels, so it would be interesting to observe how the narrative situations are mapped in such a structure. Regarding the used synthesis method, we consider that HNM-based TTS synthesis is a good approach to address storytelling speech thanks to the synthesis flexibility it allows. Nonetheless, other methods like concatenative synthesis can also be considered and compared.

7. Acknowledgements

The first author of this paper would like to acknowledge the support of the Catalan Government (SUR/ECO) for the predoctoral FI grant No. 2013FI_N 00790. We also thank Àngel Calzada and Dr. Joan Claudi Socoró for their support in the HNM synthesis implementation.

8. References

[1] D. Doukhan, A. Rilliard, S. Rosset, M. Adda-Decker, and C. d'Alessandro, "Prosodic analysis of a corpus of tales," in *Interspeech*, 2011, pp. 3129–3132.

[2] J. Adell, A. Bonafonte, and D. Escudero, "Analysis of prosodic features towards modelling of emotional and pragmatic attributes of speech," *Procesamiento del lenguaje natural*, no. 35, pp. 277–283, 2005.

[3] H. Blancafort, A. Tusón, and A. Valls, *Las Cosas del decir: manual de análisis del discurso*, ser. Ariel Letras. Editorial Ariel, 2007.

[4] O. Jokisch, H. Kruschke, and R. Hoffmann, "Prosodic reading style simulation for Text-to-Speech synthesis," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, J. Tao, T. Tan, and R. W. Picard, Eds. Springer Berlin Heidelberg, 2005, vol. 3784, pp. 426–432.

[5] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for storytelling applications," *IEEE Trans. on Audio, Speech and Language Processing*, pp. 1137–1144, 2006.

[6] C. O. Alm and R. Sproat, "Perceptions of emotions in expressive storytelling," in *Interspeech*, 2005, pp. 533–536.

[7] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, B. C., Canada, 2005, pp. 579–586.

[8] V. Francisco, P. Gervás, M. González, and C. León, "Expressive synthesis of read aloud tales," in *Artificial and Ambient Intelligence*, 2007, pp. 179–186.

[9] A. Silva, M. Vala, and A. Paiva, "The storyteller: Building a synthetic character that tells stories," in *Proc. of the Workshop Multimodal Communication and Context in Embodied Agents*, 2001, pp. 53–58.

[10] F. Burkhardt, "An affective spoken storyteller," in *Interspeech*, 2011, pp. 3305–3306.

[11] H. Buurman, "Virtual storytelling: Emotions for the narrator," Master's thesis, University of Twente, The Netherlands, 2007.

[12] A. Silva, G. Raimundo, A. Paiva, and C. Melo, "To tell or not to tell...Building an interactive virtual storyteller," in *Proc. of the Language, Speech and Gesture for Expressive Characters Symposium, Artificial Intelligence and the Simulation of Behaviour Covention*, March 2004.

[13] R. Gelin, C. d'Alessandro, O. Deroo, Q. A. Le, D. Doukhan, J.-C. Martin, C. Pelachaud, A. Rilliard, and S. Rosset, "Towards a storytelling humanoid robot," *AAAI Fall Symposium Series on Dialog with Robots*, pp. 137–138, 2010.

[14] V. A. Propp, *Morphology of the Folktale*, 2nd ed., ser. Publications of the American Folklore Society. University of Texas Press, 1968.

[15] R. Cowie, "Describing the emotional states expressed in speech," in *ISCA Workshop on Speech & Emotion*, Northern Ireland, 2000, pp. 11–18.

[16] M. Schröder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," Ph.D. dissertation, Saarland University, 2004.

[17] N. Mamede and P. Chaleira, "Character identification in children stories," in *Advances in Natural Language Processing*, ser. Lecture Notes in Computer Science, J. L. Vicedo, P. Martínez-Barco, R. Muñoz, and M. Saiz Noeda, Eds. Springer Berlin Heidelberg, 2004, vol. 3230, pp. 82–90.

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. (v.5.3.39)," retrieved 6 January 2013 from http://www.praat.org/.

[19] D. Hirst, "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding of intonation," in *Proc. of the 16th international congress of phonetic sciences*, 2007, pp. 1233–1236.

[20] J. P. Goldman, "EasyAlign: An automatic phonetic alignment tool under Praat." in *Interspeech*, 2011, pp. 3233–3236.

[21] I. Iriondo, F. Alías, J. Melenchón, and M. A. Llorca, "Modeling and synthesizing emotional speech for Catalan Text-to-Speech synthesis," in *Tutorial and Research Workshop on Affective Dialog Systems*, 2004, pp. 197–208.

[22] F. Burkhardt and W. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *Proc. of the ISCA Workshop on Speech and Emotion*, 2000, pp. 151–156.

[23] I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. M. Blanco, D. Bernadas, J. M. Oliver, D. Tena, and L. Longhi, "Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques," in *Proc. of the ISCA Workshop on Speech and Emotion*, 2000, pp. 161–166.

[24] M. Kienast, A. Paeschke, and W. F. Sendlmeier, "Articulatory reduction in emotional speech," in *EUROSPEECH*, 1999, pp. 117–120.

[25] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *J. Pers. Soc. Psychol.*, vol. 70, no. 3, pp. 614–636, Mar. 1996.

[26] À. Calzada and J. C. Socoró, "Voice quality modification using a Harmonics Plus Noise Model," *Cognitive Computation*, pp. 1–10, 2012.

[27] Y. Stylianou, "Harmonic plus Noise Models for speech, combined with statisticals methods, for speech and speaker modification," Ph.D. dissertation, École Nationale Supérieure des Télécommunications, 1996.

[28] D. Erro, "Intra-lingual and cross-lingual voice conversion using Harmonic plus Stochastic Models," Ph.D. dissertation, Technical University of Catalonia, 2008.

[29] S. Planet, I. Iriondo, E. Martínez, and J. A. Montero, "TRUE: an online testing platform for multimedia evaluation," in *Proceedings of the Second International Workshop on EMOTION: Corpora for Research on Emotion and Affect at the 6th Conference on Language Resources & Evaluation*, ser. LREC '08, Marrakech, Morocco, 2008.