Prosodic Analysis of a Corpus of Tales

David Doukhan[1]*, Albert Rilliard*[1]*, Sophie Rosset*[1]*, Martine Adda-Decker*[1,2]*, Christophe d'Alessandro*[1]

[1] LIMSI-CNRS, 91403 Orsay, France
[2] Laboratoire de Phonétique et Phonologie LPP-CNRS, UMR 7018, 75005 Paris, France
{doukhan,rilliard,rosset,madda,cda}@limsi.fr

Abstract

This paper present a prosodic analysis of a corpus of 12 tales, read by a male speaker. The work is part of a project which aims at giving a storytelling ability to a humanoid robot. One main point is to improve text-to-speech synthesis expressivity according to a semi-automatic analysis of a given tale. Automatic tagging and prosodic stylization was applied to the corpus. The extracted parameters are described and analyzed according to relevant elements of the tales structure. The results underline the expressive strategy used by the speaker to impersonate the different kinds of characters and during the different structural parts of each tale. The relevance of this prosodic parameters are then discussed in order to propose relevant instructions to enhance the expressivity of a non-uniform-units text-to-speech synthesizer.

Index Terms: Expressive prosody, Storytelling, Prosodic analysis, Corpus analysis

1. Introduction

The aim of this project is to provide the humanoid robot Nao [1] the ability to interpret tales in front of a children audience [2]. Given the raw text of a French tale, Nao will have to automatically coordinate the synthesis of entertaining gestures synchronized with a synthetic speech stream. As a constraint, it as been decided that the robot will only have one voice to read these tales and impersonate all the characters and the narrator - reproducing the situation of an adult telling a story to his/her children. The present paper analyze the prosodic variation contained in a corpus of 12 tales read by one male speaker. This prosodic analysis aims at enhancing the expressivity of the non-uniform-units text-to-speech synthesizer Acapela [3] .

Previous studies dealing with the automatic generation of synthetic storytelling speech (e.g. [4]) proposed melodic rules build to render a storyteller speaking style, or to induce the perception of an increasing climax . The rules proposed by [4] are inferred from the observation of the voices of Dutch storytellers. Such an approach is pertinent for driving a diphone synthesizer, the prosody of which is driven by a set of rules. But, if the gathered information are highly valuables for modeling some kinds of storytelling prosody, this approach cannot be directly used with a speech synthesizer like Acapela's one, selecting non-uniform units across a large corpus to concatenate them with a minimum signal processing in order to obtain the highest level of naturalness. Other works specifically design the voice of a characters: [5] presented a cepstral analysis of voices and derived a synthesis method to obtain prototypical voices for different kind of characters (the teller, a witch, a prince). For similar reasons, the description is relevant, but the method cannot be directly applied to the framework of the GVLex project.

The proposed approach extracts a set of global prosodic parameters from a large corpus of read tales, and linked them to a set of qualitative descriptors extracted from the tales' linguistic and narrative analysis. Then, the relevant prosodic variations are proposed for the different prototypical tales parts (e.g. triggering event or epilogue), different classes of characters described by gender, size, etc., or the communicative aim of the narrator. This prosodic descriptions are then used to drive the synthesizer in its units-selection task.

After presenting our textual and audio tales corpora, their linguistic and prosodic analysis is described. The prosodic parameters are then described under the light of the chosen descriptors. Then, a set of possibly relevant prosodic variations for enhancing the generation of synthetic speech in accordance with a tale's narrative structure are listed and critically discussed.

2. Corpus

2.1. Text Corpus

A corpus of 89 tales in French, representing about 50,000 words, was collected. Tales were selected so as to each present different speakers, and being suited for a 7-8 years old audience. Each tale has to be readable in about 5 minutes. A tale contains 907 word on average.

The structural and lexical elements were manually annotated, according to the following scheme. The first level of structural annotation is derived from tales' morphology [6], and presents the following parts: title, exposition, triggering event, a series of scenes (that may be interleaved with refrains), ending with the epilogue. The second structural level refers to speech turns, and associate a distinct identifier per character, identifying the narrator and the different characters. The third level refers to sentences. According to our structural hierarchy, a sentence cannot cover adjacent speech turns. Consequently, the following example *"- hé mais ! se dit-elle, ce sont des bébés !"* ("Hey, but! she told herself, it is babies!") would have been split into three sentences, each of them corresponding to a speaker turn.

The last structural level refers to passages of enumerations, such as: *"pas le blé, ni les noix ni le pain dur."* ("not wheat, nor nuts nor stale bread."), and to the elements of enumerations, which starts the lexical level of annotation (in the above example, elements of the enumeration are:<pas le blé>, <ni les noix>, <ni le pain dur>).

At the lexical level, tagging was performed for named entities (time and place), and extended named entities (nominal group and adverbial locutions). The definition of person type has been extended to identify all kinds of characters occurring in tales (humans, animals, plants, objects, ...). Finally, Part Of Speech were tagged using a LIMSI software [7].

Table 1: *Global characteristics of the 12 tales retained to constitute the audio corpus.*

	Min	Max	Mean	Total
Tale duration	215	374	299.4	3593
Word count	626	1031	805.3	9664
Vowel count	881	1304	1061.8	12741
Phon. count	1995	2914	2374.2	28491
Breath count	23	116	56.9	683
Hesitation count	1	33	11.5	138
Sentence count	50	150	80.5	966
Nb. characters	2	11	5.4	65

2.2. Audio Corpus Recording

A 12 tales subset of the text corpus has been selected for an audio corpus. The selection was aimed to promote tale diversity (Animal, Repetitive, Fairy...). The 12 tales have been told in a studio by a professional speaker, assisted by a sound engineer. The storyteller was allowed to overwrite during playback the worst portions of recording. Consequently, the final recordings contain less errors than may be found in natural speech, and is more suited to the synthesis of entertaining speech. The storyteller was allowed to change small portions of tale texts, if he thought they could not be told fluently. The recordings were digitized in high quality and downsampled at 16KHz, with a 16 bits sample size, for the analysis.

2.3. Audio Corpus Labeling

2.3.1. Labels for Phonemes, Words and Characters

The grapheme-phoneme conversion of the text was performed thanks to the LIMSI semi-automatic alignment software [8, 9] and aligned to the boundaries of word, phonemes, and extra-phonemic events (breaths, hesitations, ...). The produced phone segment boundaries tend to occur in the transition zone between two consecutive phones. However, their exact positions do generally not correspond to boundaries as manually located by human experts [10]. The automatic alignment method has the advantage of staying consistent over the whole corpus. Only major segmentation errors were hand-corrected.

Syllable boundaries were inferred from the syllabification rules described in [11]. The text structural and lexical annotations were hand corrected, to match the text that was actually said by the storyteller. A quantitative description of the corpus is provided in table 1.

A manual enrichment of text annotations was performed to provide details about the characters played by the storyteller. Character's age (coded as: kid, teenager, adult, old), their gender (male, female), and their kind (human, wolf, fairy, knife, ...) were added whenever appropriate to the tale.

2.3.2. Vowel-Based Prosodic Features

The stylisation proposed by the Prosogram [12] (version 2.7) was applied in order to measure perceived intonation. The stylisation performed by the Prosogram consists to associate to each vowel either a level tone, a glide, or several glides, according to a perceptive threshold. From this stylization were extracted: The lowest and highest pitch value within a nucleus; the inter-syllable pitch and intensity differences (calculated as the difference between the values at the end and at the beginning of two adjacent vowels); the tone's shape (0 for static tone, 1 for glide); and the maximal intensity within the nucleus.

Pitch is expressed in semitones with 1 Hz as the reference value. Intensity is measured in dB. 2% of vowels were rejected by the Prosogram, which is consistent with the analysis described by [13] on a French corpus. Those vowels were thus ignored in the reported measures of pitch and energy.

3. Storytelling Prosody

3.1. Global measures

A general analysis of the whole corpus was done, comparing such mean values of the extracted prosodic parameters to the same values found in the literature for others styles of French speech.

[13] presented a work on 20 French newslike utterances. While [13] observed glide tones on 2% of the stylized vowels, 13% of the vowels in the tale corpus were assigned glide tones, accounting for important melodic dynamism of storytelling intonation.

Table 2: *Comparison of speaking styles presented by [14] (News: Radio News, Pol: Political Address, Conv: Conversation) and the Storytelling style contained in the complete corpus (StT) or restricted to the narrator speaking turn (Nar.). See text for prosodic parameters.*

Descriptor	News	Pol	Conv	StT	Nar.
SR	5.8	4.8	5.3	6.15	6.2
PT	10.97	31.67	16.73	29.55	24.0
NS	15	8	16	7.4	7.4
PR	10.5	10.5	7.4	17.4	16.1

[14] describe three speaking styles: Radio News, Political Address and Conversational speech, on the basis of 10 minutes of speech per style. Table 2 compares mean prosodic features of storytelling style prosody to these three styles. For the storytelling style, mean values are presented for the whole corpus and for the passages where the narrator speaks, in order to have a more constraint style, without variations due to impersonation of a given character. The measured parameters are : the speech rate (**SR**) measured as the number of syllables by seconds (pauses excluded); the pausing time percentage (**PT**); the number of syllables (**NS**) between two pauses; and the pitch range (**PR**) between the 0.05 and the 0.95 quantiles pitch distribution.

The reported observations show rhythmic similarities (pause rate and number of syllables between pauses) between the Storytelling and the Political styles. The pitch range is clearly wider for storytelling than for all the other reported styles - about 6 semitones more. No important differences were found between the complete tale corpus and its narrator subset - which represent the main part of the corpus in terms of speech time.

3.2. Prosodic variation and tales' narrative structure

The stylized prosodic parameters extracted from Prosogram were averaged on the different sequences of the corpus to extract the following descriptors of the prosodic variations observed:

- The mean pitch (**MP**) and the mean intensity (**MI**) expressed on a given sequence.
- The mean inter-syllable difference for pitch (**PD**) and intensity (**ID**).

Table 3: *Prosodic comparison of storyteller speech given the structural part of the tale (see section 2.1). Prosodic descriptor used are defined in sections 3.1 & 3.2, NV being the number of vowels contained in these structural sequences.*

	title	exp	trig evt	scene	refrain	epilogue
NV	75	1376	484	5115	473	536
SR	5.4	6.3	6.4	6.2	6.3	6.1
PT	5.0	24.9	21.6	20.1	15.0	20.7
MP	85.1	84.6	83.4	84.2	84.1	83.4
PD	4.9	4.1	4.5	3.1	2.8	3.2
PR	21.0	17.1	16.6	15.7	15.4	15.0
GL	27.0	16.1	17.2	11.3	12.2	12.1
MI	67.2	66.3	64.8	66.1	66.7	64.6
ID	2.7	2.9	3.1	2.8	2.9	2.6
IR	13.3	14.9	14.6	15.1	15.0	11.8

- The pitch range (**PR**) and intensity range (**IR**) measured as the difference between the .95 and .05 quantiles of each distribution.
- The percentage of vowels stylized as glides(**GL**).

Table 3 presents the storyteller's prosodic characteristics, for the different narrative structures of each tale (as presented in section 2.1). Again, to avoid a bias linked with the impersonation of characters, the values reported here only concern the narrator's speech turns. Note that the pausing time percentage values reported here are smaller than those reported in table 2, since pauses between narrative structures were not taken into account.

Analysis of variance were done on the reported measures' distributions, with post-hoc Tukey HSD test to compare all sequences mean value. The scenes sequences, accounting for 63.5% of the syllables, present values around the average for all prosodic descriptors.

Title enunciation was clearly distinguishable from other structural parts, having the highest glide rate, a high mean intensity and inter-syllable pitch difference (ranked similar to those of the exposition and of the triggering event but significantly higher than for other sequences), the larger pitch range, and the slowest speech rate. Note that its low pausing time percentage is irrelevant, since titles are based on short passages.

The beginning of tales (i.e. the exposition and the triggering event sequences) share prosodic similarities (high glide rate and inter-syllable pitch difference), accounting for the important melodic dynamism which may be aimed at capturing attention. They significantly differ on mean intensity, lower for the triggering event, which may be linked to the use of a soft voice inducing intimacy and suspense.

Refrains have the smallest pausing time percentage observed on the tale corpus, and the smallest inter-syllable mean pitch difference, comparable to the scenes and the epilogue. Epilogue was mainly characterized as having the lowest pitch and intensity ranges, and a low speech rate. Moreover it has the lowest mean pitch and mean intensity with the triggering event; and is amongst the lowest cluster for inter-syllable pitch and intensity differences with refrains and scenes. All these cues account for quiet and flat intonation at the end of tales.

3.3. Prosodic variation and characters impersonation

For each character of the 12 tales, the same mean prosodic descriptors than above have been extracted, in order to sort out the most pertinent ones and the prosodic variations used to impersonate different characters. The percentage of unstylized vowels (**UV**), calculated as the percentage of vowel considered as devoiced by the Prosogram, was also added to the measures. In order to have reliable data, only the main characters were kept, for which sufficient data has been recorded (i.e. if they have pronounced at least 50 syllables). 32 characters, plus the narrator correspond to this criterion. The narrator has the most varying prosodic patterns, that can easily be accounted for due to its over-representation in the corpus and to the different narrative situations he has to render. He will here be used as a reference for the voice of the storyteller, allowing to rate the most important divergences used to impersonate distinctive characters.

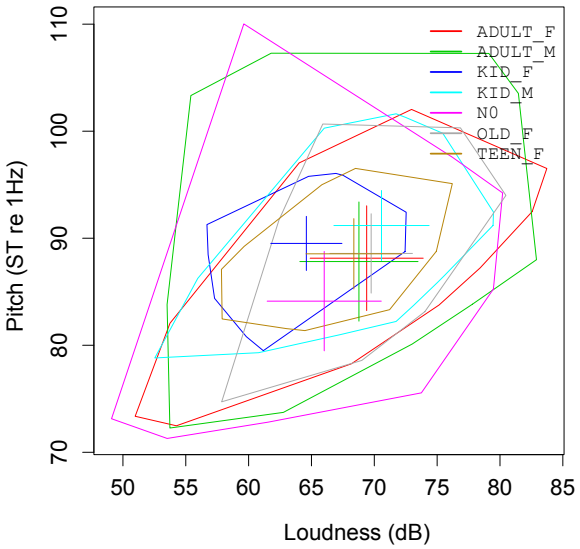

Figure 1: *Pitch and Intensity distributions for the narrator (N0) and different categories of embodied characters varied for age (kid, teenager, adult, old) and gender (M/F)*

The distribution of the pitch and intensity values measured for all characters, grouped by gender and by age is displayed in figure 1. It shows that the storyteller tends to exhibit a higher pitch and intensity when impersonating a character than for the narrator's passage. The youngest characters are linked with a higher pitch register than other characters. Kids have the highest pitch and intensity register, while girls had the lowest intensity register. Mean pitch and intensity for adult male and female characters are comparable.

For a more detailed analysis, these characters have been separated in two sets according to their gender. Then, a Principal Component Analysis (PCA) was run on each set. Two birds characters which does not have a clear gender were removed from the characters. Three more characters have also been removed due to their very peculiar prosodic parameters, that gathered most of the PCA variance, mainly due to a high percentage of devoiced vowels. As the reference voice, the narrator has been added to both sets. The female set regroups 14 characters (including the narrator), such as "mother mouse", fairies, a hen, a mother, a grand mother or little girls. The male set regroups 15 characters (including the narrator), such as a bear (and many other animals), a farmer, a deaf man, an adult hero, a boy or a small monster.

For the female set, the first principal component is linked with the measures of local pitch and intensity variations (such as glides and inter-syllable pitch difference) and negatively with the percentage of devoiced vowels. The second component is linked with the mean pitch and the number of syllables without pause. The female characters, plotted over the two first principal components, may be regrouped into four clusters (thanks to a hierarchical clustering on the same prosodic parameters) that have pertinent narrative characteristics: (1) two young girls that have a leading role (quite high and flat pitch, high percentage of devoiced vowels), (2) adult females having a supporting role with the narrator (hence close to narrator prosodic characteristics), (3) older females or characters in charge of an important responsibility - e.g. "mother mouse" who has to care children (higher mean intensity, more local pitch variations), and (4) one single supporting character acting with authority to help the hero, and played with the highest mean pitch and intensity.

The male set show a slightly more complex figure, with the first principal component opposing the mean pitch to measures of inter-syllable pitch differences and glides. The second component opposes the pitch range to the percentage of unvoiced vowels and inter-syllable intensity differences. The male characters, plotted over these two principal components, may be clustered into five clusters that also have pertinent narrative characteristics: (1) young and unexperimented (threatened) leading roles (highest mean pitch and intensity with few local variations), (2) supporting roles (high mean pitch) - these characters are quite close to the preceding cluster, (3) supporting role impersonating old men and/or characters having a great presence with the narrator (medium pitch with a notable local variation) (4) supporting role helping or counselling the hero (medium pitch with a high pitch range and pitch dynamic) (5) a single character : a big bear played with the lowest pitch, high local changes in intensity and more devoiced vowels than other characters.

4. Conclusions and Future work

We have reported on the recording, labelling and prosodic analysis of a one hour storytelling corpus recorded by one male French speaker. From this large and coherent data, it is clear that storytelling induce much more variations than most of the other speaking styles (e.g. political address, radio news), and constitute a challenge for speech synthesizers. The narrative structure of tales account for interesting prosodic variations, that are linked for example to the expression a climax during the triggering event (see [4]), or to a calm final for the epilogue. But the most striking prosodic variations are performed during the impersonation of the tales's characters. PCA and hierarchical clustering performed on 27 different characters, labeled for gender, age and role in the tale, allowed an interesting classification of prosodic parameters linked to different character types; 4 character types have been found for female characters, and 5 for male characters.

The prosodic parameters used - either mean and range of pitch and intensity, or percentage of glide tones - may possibly be transformed into rules to drive a non-uniform-units text-to-speech synthesizer.

Future work will consists in considering bigger sets of prosodic features, including prominences [14] and voice quality parameters. Most prosodic descriptors considered in this study were static, and provided limited informations on the dynamics within categories, which will be investigated in the next studies. While information concerning speakers global characteristics have been made available, they may be refined by adding affect and emotional tags at the sentence level. Affect and emotion tagging may be reductive, and [15] recommended to focus rather of identifying prototypical pitch contour. Both approachs seems complementary and will be investigated.

5. Acknowledgements

This work has been funded by the French project GV-LEx (ANR-08-CORD-024 http://www.gvlex.com).

6. References

[1] Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J. and Maisonnier, B., "Mechatronic design of NAO humanoid", in Proceedings of the 2009 IEEE international conference on Robotics and Automation, ICRA'09, 2124–2129, IEEE Press, 2009.

[2] Gelin, R., d'Alessandro, C., Le, Q. A., Deroo, O., Doukhan, D., Martin, J.-C., Pelachaud, C., Rilliard, A. and Rosset, S., "Towards a Storytelling Humanoid Robot", in AAAI Fall Symposium Series on Dialog with Robots, 137–138, 2010.

[3] http://www.acapela-group.com/.

[4] Theune, M., Meijs, K., Heylen, D. and Ordelman, R., "Generating expressive speech for storytelling applications", IEEE Transactions on Audio, Speech, and Language Processing, 14(4):1137–1144, 2006.

[5] Přibil, J. and Přibilová, A., "Application of Expressive Speech in TTS System with Cepstral Description", in A. Esposito, N. G. Bourbakis, N. Avouris and I. Hatzilygeroudis, eds., Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction, 200–212, Springer-Verlag, Berlin, Heidelberg, 2008.

[6] Propp, V., Morphology of the Folktale, University of Texas Press, 1968 (orig 1928).

[7] Allauzen, A. and Bonneau-Maynard, H., "Training and evaluation of pos taggers on the french multitag corpus", Proceedings of the Sixth International Language Resources and Evaluation (LREC08), Marrakech, Morocco, 2008.

[8] Adda-Decker, M. and Lamel, L., "Pronunciation variants across system configuration, language and speaking style", Speech Communication, 29(2-4):83–98, 1999.

[9] Gauvain, J., Adda, G., Adda-Decker, M., Allauzen, A., Gendner, V., Lamel, L. and Schwenk, H., "Where Are We in Transcribing French Broadcast News?", in Ninth European Conference on Speech Communication and Technology, ISCA, 2005.

[10] Astesano, C., Bertrand, R., Brousseau, M., Chafcouloff, M., Di Cristo, A., Ghio, A., Hirst, D., Lapierre, S., Nicolas, P., Roméas, P. et al., "The PACOMUST Project, a corpus of multistyle continue speech: objectives and methodological choices", Travaux de l'institut de Phonétique d'Aix, 16:9–38, 1995.

[11] Adda-Decker, M., Boula de Mareuil, P., Adda, G. and Lamel, L., "Investigating syllabic structures and their variation in spontaneous French", Speech Communication, 46(2):119–139, 2005.

[12] Mertens, P., "The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model", in Proceedings of Speech Prosody, 23–26, 2004.

[13] Patel, A., Iversen, J. and Rosenberg, J., "Comparing the rhythm and melody of speech and music: The case of British English and French", The Journal of the Acoustical Society of America, 119:3034–3047, 2006.

[14] Roekhaut, S., Goldman, J. and Simon, A., "A Model for Varying Speaking Style in TTS systems", in Fifth International Conference on Speech Prosody, Chicago, IL, 2010.

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