Rhythmic Patterns and Literary Genres in Synthesized Speech

Elisabeth Delais-Roussarie[1], *Damien Lolive*[2], *Hiyon Yoo*[1] and *David Guennec*[2]

[1] UMR 7110-LLF & Université Paris-Diderot, France
[2] IRISA, ENSSAT, Université Rennes 1, France

Elisabeth.roussarie@wanadoo.fr, yoo@linguist.univ-paris-diderot.fr,
{damien.lolive/david.guennec }@irisa.fr

Abstract

In this paper, the rhythmic patterns observed in natural and synthesized speech are compared for three literary forms (rhymes, poems, and fairy tales). The aim of the comparison is to evaluate how rhythm could be improved in synthesized speech, which could allow adapting it to specific styles or genres.

The study is based on the analysis of a corpus of six rhymes, four poems and two extracts from fairy tales. All texts were recorded by three speakers and were generated with two distinct synthesized voices. The comparison of the rhythmic patterns observed is done by analyzing duration in relation to prosodic structure in the various data sets. This approach allows showing that rhythmic differences between synthesized and natural speech are mostly due to the marking of prosodic structure.

Index Terms: Rhythmic patterns, phono-genre, speech synthesis, prosodic structure.

1. Introduction

In the last twenty years, the overall quality of synthesized speech has greatly improved with the emergence of new TTS techniques, including corpus-based concatenative speech synthesis systems ([1] and [2]). Nevertheless, generating a natural-sounding prosody remains a challenge (see [3] among others). More specifically, the rhythmic component of these systems often sounds odd and unnatural, and needs to be improved for using synthesis in a wide range of applications (games, educational software, etc.).

In a research project aiming at using speech synthesis to teach writing skills to primary school pupils, it appeared important to improve the TTS system to allow it reading more accurately different types of data: fairy tales, poetry and rhymes. In order to achieve such a goal, a comparison of the rhythmic patterns obtained in natural and synthesized speech in the different genres was achieved. Regarding synthesized speech, one of our hypotheses was that more accurate rhythmic patterns would be observed in fairy tales, since the corpora used to select the speech units for the TTS system are mostly composed by read sentences extracted from audio-books. Since our findings do not really confirm the hypothesis, it seemed to us important to understand the reasons why the rhythmic patterns were more accurate for poems and rhymes than for fairy tales.

This paper is organized as follows. Section 2 provides a description of the data and methods used for the study. In section 3, the results obtained from the prosodic analyses by comparing duration patterns and speech rate along at least two dimensions (synthesized vs. natural speech, difference among the literary genres) are presented. Main findings are discussed in section 4 by focusing on what is crucial to improve TTS systems.

2. Corpus and Methodology

2.1. Corpus

The corpus used to study the rhythmic patterns obtained in natural and synthesized speech consisted of three distinct types of texts that could all be addressed to children: six rhymes, four poems and two extracts from fairy tales. Table 1 summarizes the exact composition of the corpus according to literary genres. Differences among speakers and synthesized voices results mostly from schwa insertion or deletion, and from word omission (in the pronunciation of titles for poems and rhymes for instance). The effective number of syllables obtained for the three speakers and the two synthesis systems is given in the last column.

Table 1. *Corpus composition*

Speaking Style	Number of words	Number of syll.	Effective number of syllables (natural *vs.* synth)
Rhymes	158	228	683 syll. / 454 syll.
Poems	290	422	1347 syll. / 808 syll.
Tales	522	777	2323 syll. / 1538 syll.
Total	970	1427	4353 syll. / 2800 syll.

The set of texts was recorded by three speakers (two males and one female) in a sound-proof room. Time for reading and rehearsing texts was given to the participants before recording. Among the three speakers, two were reading the texts as parents would read a story to their children, whereas the third one is a trained actor and was reading the texts with great expressivity.

As for the synthesized stimuli, they were produced by a corpus-based TTS system as presented in [4] and for which pre-selection filters are used instead of a target cost. For the purpose of this study, the ordered filters set we used is the following:

1. Unit label (cannot be relaxed).
2. Is the unit a Non Speech Sound (cannot be relaxed) ?
3. Is the phone in the last syllable of its sentence ?
4. Is the phone in the last syllable of its major prosodic group (IP) ?

5. Is the current syllable in word end ?

6. Is the current syllable with a rising intonation ?

During the best unit sequence search, if the number of units corresponding to a given set of filters is too low, the last filter of the set is relaxed. By reducing the number of applied constraints, the search space becomes wider. In any case, the two first filters are kept. Furthermore, a penalty is applied to phoneme classes for which concatenation seems to be risky (see [5]). Indeed, we consider that joining two units on a vowel is more likely to produce an artefact than when joining is made on the silent part of a plosive or even with a fricative. Concerning prosody, no specific treatment is made, and the only constraints that may improve the generated speech rhythm are the pre-selection filters, as they impose positional constraints to selected units. Finally, pauses are placed at designated places by the system: a pause is for instance inserted after each punctuation mark. Note also that their duration remains fixed, and is not related to the length of the preceding speech stretch.

For this study, two distinct synthesized voices were used. They differ according to the way they were produced:

- voice SY-P, a male voice, is based on a corpus of 10 hours extracted from an audiobook, i.e. a novel read by an actor.
- voice SY-A, a female voice, consists of 7 hours of read speech, the read items being specifically designed to build up a speech synthesis system.

The differences in the content and the size of the corpora lead to consider voice SY-P as more expressive than voice SY-A, which is more neutral.

To generate the synthesized stimuli, the structure in stanza and lines for poems and rhymes was represented by using punctuation marks such as comma. The three stanza in (1), which are extracted from a poem (*La fourmi*, R. Desnos), were typed as shown in (2) to obtain the synthesized version.

(1) *Une fourmi traînant un char*
plein de pingouins et de canards
ça n'existe pas, ça n'existe pas

Une fourmi parlant français
parlant latin et javanais
ça n'existe pas, ça n'existe pas

eh ! et pourquoi pas !

(2) Une fourmi traînant un char, plein de pingouins et de canards, ça n'existe pas, ça n'existe pas. Une fourmi parlant français, parlant latin et javanais, ça n'existe pas, ça n'existe pas. Eh ! Et pourquoi pas !

As shown above, the end of stanzas is always encoded by a full stop, when no punctuation mark was used in the original text. The lines were encoded by a comma, except when ending with a punctuation mark in the text. The other parts of the text remain unchanged.

2.2. Methodology

The data were first orthographically transcribed and segmented in utterances using Praat [6]. The orthographic transcription was then phonetized, and the audio signal automatically segmented into phones, syllables, and graphemic words by means of the speech processing script EasyAlign [7]. The obtained phonetic transcriptions and acoustic segmentations were controlled and corrected when necessary. The entire annotated data set was then used to carry out the rhythmic and prosodic analysis.

To generate the duration patterns and to analyze and compare pause durations and speech rates according to speakers and genres, vowels were chosen as the base unit instead of syllables. This choice results from the fact that syllable structures vary a lot in French and syllabic duration cannot be a robust indicator to evaluate the lengthening rate. As the number of vowels located in the different prosodic positions was limited because of the size of the corpus, it was difficult to normalize duration. We thus decided to make, a distinction between long and short vowels, even if such a distinction does not exist in the French phonological system. Nasal vowels ([ɔ̃],[ɑ̃], [ɛ̃] and [œ̃]) and sequences composed of a semi-vowel and a vowel in nuclear positions (as, for instance, [jɛ̃] in *tiens* [tjɛ̃], [wa] in *noir* [nwaʁ]) were thus encoded as long vowel, whereas the remaining oral vowels were considered as short.

Since previous studies on French prosody showed that phrasing, intonation and accentuation are highly intertwined in this language (e.g. among others [8]), all sentences from the different texts were segmented in prosodic phrases, a distinction being made between three levels of phrasing (prosodic word PWD, phonological phrase PP and intonational phrase IP). Rules were used to derive the prosodic phrases from the text, i.e. from the morpho-syntactic structure and the number of syllables (see, among others, [9], [10] and [11]). Such an approach has the advantage of avoiding a certain circularity.

Since the last syllable of prosodic phrases is considered as accented in French and is usually lengthened (see [12]), we distinguish three categories of accented syllables to compare the lengthening rate of the accented syllables in relation to their prosodic position:

- AC-PWD, which corresponds to the last metrical syllable of a prosodic word, i.e. a word from a lexical category such as Verb, Noun, Adjective and Adverb (see [13] and [14] among others);
- AC-PP, which coincides with the last metrical syllable of a minor phrase, i.e. of the lexical head of a syntactic projection (see [9], [15] and [16] among others);
- AC-IP, which corresponds to the last metrical syllable of any IP, IP boundaries being located at the end of a clause, a detached syntactic constituent, or a line (in poems and rhymes), see [14], [17] and [18] among others.

3. Results

Duration patterns obtained for natural and synthesized speech allows analyzing and comparing speech rates, pause duration and distribution, and prosodic structure marking. The results are presented in the two following sub-sections.

3.1. Speech rate and pausing

The total duration of the various readings was used to calculate for each speaker and each genre the speech and articulation rates as well as pause durations. The difference between articulation and speech rates relies on the fact that pauses are not taken into account to calculate articulation rate (see [19]). Table 2 summarizes the results obtained for each speaker and in the three distinct genres. The first two rows indicate respectively speech and articulation rates in number of phones by second, whereas the last two rows are of interest to study the duration and distribution of pauses.

Table 2. *Speech and articulation rates in phones/sec, and pause duration and percentage of pauses (related to the total duration of readings)*

Rhymes	LOD	DRE	GOR	SY-A	SY-P
Average speech rate (ph./sec.)	9.9	7.35	7.08	7.63	9.09
Average articulation rate (ph./sec)	12.09	7.83	8.53	9.79	12.61
Total pause duration (ms)	2178.92	1449.22	2573.76	3025	3000
Average % of pauses	25.27	13.60	24.15	29.06	33.80
Poems	LOD	DRE	GOR	SY-A	SY-P
Average speech rate (ph./sec.)	10.6	8.16	6.28	8.26	9.45
Average articulation rate (ph./sec)	13.60	9.32	8.72	10.70	12.85
Total pause duration	1534	1373.36	2590.52	2000	2000
Average % of pauses	27.38	18.10	33.85	28.17	31.29
Tales	LOD	DRE	GOR	SY-A	SY-P
Average speech rate (ph./sec.)	10.58	8.74	8.18	9.31	10.79
Average articulation rate (ph./sec)	14.99	10.08	11.09	11.36	13.68
Total pause duration	1331.33	763.40	1482.06	992	992.14
Average % of pauses	32.82	18.06	29.96	21.96	24.79

The articulation and speech rates observed for each genre vary a lot, but one cannot say that synthesized voices differ from natural one. LOD and SY-P speak faster than the other speakers in all genres, whereas GOR and DRE obtain the lowest rates. Within a given literary genre, the speech and articulation rates obtained by synthesized voices are included in the variation space derived from the three natural voices.

A comparison across genres shows that human speakers adapt their speech and articulation rates to genres, slower rates being used for rhymes and poetry reading, whereas this adaptation is less clear for synthesized speech. This derives from the fact that the same corpus and the same unit selection procedure are used in all genres by the two synthesized voices. Differences are however minor.

Concerning pausing, there is an important difference between natural and synthesized speech, across genres as well as in general. The pause proportion is lower in rhymes than in tales in the productions of all three speakers. By contrast, there is a higher proportion of pause in rhymes than in tales for the two synthesized voices. In addition, pause duration appears to be related to articulation rate in natural voices, longer pauses being observed in rhymes and poetry.

By and large, no great difference is to be observed between natural and synthesized speech concerning speech and articulation rates. Indeed, rates vary a lot between speakers, but synthesized voices vary along the same lines. By contrast, pause duration and proportion differ between synthesized and natural voices. Since pauses may also be used to encode prosodic structure, a careful analysis of duration patterns with respect to prosodic structure is provided in the following section. Finally, the main difference between the two synthesized voices comes from their nature as SY-A is read speech while SY-P is more expressive. For instance, the average speech rates and articulation rates are very different for both voices.

3.2. Prosodic structure and duration patterns

In French, syllabic and vocalic lengthening mostly indicates phrasing and accentuation. Indeed, accented syllables, which correspond to the last full syllable at any level of prosodic structure, are lengthened, the lengthening rate being generally related to the level of phrasing (e.g. [15], [18]). Lengthening rates were thus computed by comparing the duration of vowels in unaccented syllables with the duration of the nucleus of any last metrical syllables (i.e. accented syllables) at the level of the prosodic word (PWD), the phonological phrase (PP) and the intonation phrase (IP). Table 3 summarizes the results obtained per genre. Mean duration of vowels in unaccented syllables is given in the first line of each genre, the lengthening rate being given in the following lines.

As shown in table 3, there is a relatively important variation in the duration of vowels in unaccented positions between the different genres, especially for human speakers. In general, vowels in unaccented positions are longer in rhymes and poems than in tales. By contrast, no clear variation is observed between genres for synthesized voices. This result confirms the fact that human speakers adapt their speaking rate according to genres, in contradistinction to synthesized voices.

As far as edge marking is concerned, lengthening always occurs at the end of the three distinct levels of phrasing, i.e. prosodic word, phonological phrase and intonational phrase, in synthesized as well as in natural speech. Across all genres, lengthening rate is from 10 to 20 %, at PWD level, 30 to 60% at PP level, and 80 to 180% at the IP level. These rates correspond to what is often said about French durational patterns. In rhymes and to a lesser extend in poetry, lengthening rates do not clearly allow distinguishing the three distinct levels of phrasing (e.g. differences between PWD and PP for LOD, DRE and SY-P in rhymes, and differences between PP and IP for LOD and GOR in poetry). Note also that lengthening rates marking IP boundaries are more important in all genres for SY-A than for human speakers; in the case of SY-P, it is proportionally more important in tales.

Table 3. *Mean duration of vowels in unaccented syll. (in ms) and lengthening rate (in %) at three levels of phrasing (PWD, PP and IP)*

Rhymes	LOD	DRE	GOR	SY-A	SY-P
Mean Unacc. duration	66	130	93	81	68
Length. Rate AC-PWD	30%	20%	20%	20%	30%
Length. Rate AC-PP	20%	20%	50%	30%	10%
Length. Rate AC-IP	90%	70%	70%	150%	60%
Poems	LOD	DRE	GOR	SY-A	SY-P
Mean Unacc. duration	67	110	95	78	69
Length. Rate AC-PWD	10%	20%	40%	20%	20%
Length. Rate AC-PP	60%	40%	100%	50%	40%
Length. Rate AC-IP	60%	70%	80%	190%	80%
Tales	LOD	DRE	GOR	SY-A	SY-P
Mean Unacc. duration	59	99	78	77	65
Length. Rate AC-PWD	10%	20%	10%	20%	20%
Length. Rate AC-PP	20%	40%	50%	40%	30%
Length. Rate AC-IP	80%	80%	100%	190%	100%

On the whole, the duration patterns obtained for synthesized speech in all genres are relatively comparable to what is observed in natural speech: the different levels of phrasing are encoded by a lengthening, whose relative rate varies in relation to boundary strength (see [15], [17] and [20] among others).

4. Discussion

The comparison between synthesized speech and natural speech does not show strong differences. The variation that occurs in speech and articulation rates does not allow distinguishing natural speech from synthesized one. Concerning final lengthening and edge marking, the prosodic analysis clearly showed that final lengthening occurs in natural and in synthesized speech, despite some differences in the lengthening rate observed at the level of the IP for SY-A (in all genres) and for SY-P (in tales, to a lesser extend). It is doubtful however that these differences explain the lack of naturalness in rhythm. By listening to synthesized stimuli we were surprised by the quality of the rhythmic patterns observed in rhymes, especially for SY-A. Indeed, they sounded very natural in comparison to those obtained for tales. So, the encountered problems in rhythm cannot be attributed to extra-lengthening at IP level. Since speech and articulation rates on the one hand, and durational marking of the prosodic structure, on the other, cannot be invoked to account for the lack of naturalness in the rhythmic patterns, other explanations have to be found. In fact, two lines of research are worth exploring. Firstly, no correlation between speech rates, boundary strength and pause duration is observed in synthesized speech, whereas such a correlation exists in natural speech. Indeed, prosodic phrases such as PPs and IPs tend to have the same number of syllables or the same duration in French (see, among others, [9], [10], [11] and [21]), and pause durations may be of importance to obtain isochrony. In synthesis speech, pause duration remain constant. In addition, tonal patterns probably play a role in the development of rhythmic patterns. By inserting a comma at the end of each line, the realization of a non-final melodic contour (i.e. continuation rise) was forced in rhymes and poetry. Since this contour was repeated regularly, it reinforced the impression of rhythm. By contrast, the form and the occurrence of tonal movements are less controlled in tales. As a consequence, the recurrence of prosodic patterns, which is crucial for rhythm, was not obtained.

5. Conclusion and perspectives

The analysis of the duration patterns observed in natural and synthesized speech for the three literary genres showed clearly that duration cannot by itself explain the lack of naturalness of the rhythmic patterns in speech synthesis. Values obtained for segmental duration and edge marking are indeed comparable in all cases. Further research on a larger corpus is necessary. In addition, three points are forth investigating to improve the unit selection procedure in the speech synthesis system and, henceforth, rhythmic patterns:

- Clearly distinguishing the various levels of phrasing: at present, lengthening rates observed at the end of the three levels of phrasing may lead to treat PWD and PP, on the one hand, and IP on the other. In natural speech, rates are located along a continuum, in all genres and for all speakers;
- Taking into account the form of the tonal movements realized on accented syllables: the procedure used to generate the synthesized stimuli forced to insert a specific tonal contour at the end of each line, i.e. at a reasonable distance in terms of number of syllables;
- Adapting lengthening rates, articulation rates and pause duration to genres, but also to satisfy some kinds of correlation.

6. Acknowledgements

The research presented here is supported by the French Investissements d’Avenir - Labex *EFL* program (ANR-10-LABX-0083). It is also partly achieved within the ANR research project *SynPaFlex* (ANR-15-CE23-0015).

7. References

[1] Y. Sagisaka, “Speech synthesis by rule using an optimal selection of non-uniform synthesis units”, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp.679-682, 1988.

[2] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 373-376, 1996.

[3] M. Schröder, “Expressive Speech Synthesis: Past, Present, and Possible Futures,” in *Affective Information Processing*, pp. 111-126, London: Springer, 2009.

[4] D. Guennec, D. and D. Lolive, “Unit Selection Cost Function Exploration Using an A* Based Text-to-Speech System”, in P. Sojka, A. Horák, I. Kopeček and K Pala, *Text, Speech and Dialogue 2014*, LNCS, Springer, Heidelberg, vol. 8655, pp. 432–440, 2014

[5] P. Alain, J. Chevelu, D. Guennec, G. Lecorvé and D. Lolive “The IRISA Text-To-Speech System for the Blizzard Challenge 2015”, Blizzard Challenge 2015 Workshop, 2015.

[6] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (Version 5.5)”. www.praat.org, 2014.

[7] J.-Ph. Goldman, “EasyAlign: an automatic phonetic alignment tool under Praat”, *Proceedings of Interspeech 2011*, pp. 3233-3236, 2011.

[8] B. Post, “The multi-faceted relation between phrasing and intonation in French”, in C. Gabriel & C. Lleó, *Intonational Phrasing at the Interfaces: Cross-Linguistic and Bilingual Studies in Romance and Germanic*, pp. 44-74, Amsterdam: Benjamins, 2011.

[9] E. Delais-Roussarie, “Phonological phrasing and accentuation in French”, in M. Nespor and N. Smith (eds), *Dam phonology: HIL phonology papers II*, den Haag: Holland Academic Graphics, pp. 1-38, 1996.

[10] P. Martin, “Prosodic and rhythmic structures in French”, *Linguistics* 25, pp. 925-949, 1987

[11] V. Pasdeloup, “A prosodic Model for French Text-to-speech synthesis : A psycholinguistic approach”, in G. Bailly, C. Benoit and T.R Sawallis (eds), *Talking Machines: Theories, Models, and Designs*, Elsevier Science Publishers, pp. 335-48, 1992.

[12] J. Fletcher, “ Rhythm and final lengthening in French”, *Journal of Phonetics* 19, pp. 193-212, 1991

[13] P. Mertens , J.-P. Goldman, E. Wehrli and A. Gaudinat, “La synthèse de l’intonation à partir de structures syntaxiques riches”, *Traitement Automatique des Langues* 42 :(1), pp. 142-195, 2001.

[14] M. Nespor and I. Vogel, Irene, *Prosodic phonology*, Dordrecht: Foris, 1986.

[15] B. Post, *Tonal and phrasal structures in French intonation*, Den Haag: Holland Academic Graphics, 2000.

[16] E. Selkirk, “On derived domains in sentence phonology”, *Phonology Yearbook 3*, pp. 371-405, 1986.

[17] E. Delais-Roussarie, B. Post, M. Avanzi, C. Buthke, A. Di Cristo, I. Feldhausen, S-A. Jun, P. Martin, T. Meisenburg, A. Rialland, R. Sichel-Bazin, and H. Yoo, “Intonational Phonology of French. Developing a ToBi system for French”, in S. Frota and P. Prieto (eds), *Intonation in Romance*, Oxford University Press, pp. 63-100, 2015.

[18] C. Portes and R. Bertrand, “Permanence et variation des unités prosodiques dans le discours et l’interaction”, *Journal of French Language Studies* 21, pp. 97-110, 2011.

[19] A-C. Simon, A. Auchlin, M. Avanzi and J.-Ph. Goldman, “Les phonostyles: une description prosodique des styles de parole en français”, in M. Abécassis and G. Ledegen (eds), *Les voix des Français : en parlant, en écrivant*, Bern : Lang, pp. 71-88, 2010.

[20] E. Delais-Roussarie and I. Feldhausen, “Variation in Prosodic Boundary Strength: a study on dislocated XPs in French”, in: N. Campbell, D. Gibbon and D Hirst (eds), *Proceedings of Speech Prosody 2014,* Dublin, May 2014, pp. 1052–1056, 2014.

[21] F. Wioland, *Prononcer les mots du français. Des sons et des rythmes*, Paris: Hachette, 1991.