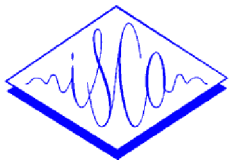

7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

Adding Speaking Style to a TTS system

Jean-Philippe Goldman[1,2], *Sophie Roekhaut*[3,4], *Anne Catherine Simon*[2]

[1] Department of Linguistics, University of Geneva, Switzerland;
[2] Institut Langage & Communication/Valibel – Discours & Variation,
Université catholique de Louvain, Belgium
[3] TCTS Lab, University of Mons - UMONS, Belgium;
[4] CENTAL, Université catholique de Louvain, Belgium ;

jean-philippe.goldman@unige.ch,
sophie.roekhaut@umons.ac.be,
anne-catherine.simon@uclouvain.be

Abstract

This paper aims to enhance the performance of a TTS system by generating various speaking styles. First we describe three speaking styles (Radio News, Political Address and Conversation) and compare the prosodic features found in these authentic styles with the prosody in "neutral" speech uttered by the eLite TTS system ([1]). Differences concern about 20 prosodic characteristics (F0 span, speech rate, pauses and hesitation, primary and secondary accentuation, schwa deletion, etc.). In order to make the neutral speech similar to a typical speaking style, prosodic characteristics are implemented within the TTS system itself or during a post-processing step. The quality of the "stylized" synthesis is evaluated by comparing it to the original style.

Index Terms: speaking styles, speech synthesis, French prosody, accentuation, pauses, hesitations.

1. Introduction

The goal of this study is to describe the prosody of several speaking styles in order to design new parameters for the fine-tuning of a non-uniform-unit-selection (NUU) text-to-speech system (TTS). Prosodic modelization of each style is based on automatic measurements taken from 3 samples of each style. The enhanced system allows a wider variety of speaking styles from a single voice database.

It is well known that a TTS system selects units from a database by taking into account the resulting continuous intonation and spectral properties of the signal. It does not allow a fine-tuning of the synthesized prosody, which would be hazardous in any respect, given the risk of reduced naturalness.

The very notion of speaking style is a nebulous one [2]. Speaking styles are supposedly recognizable, or salient, manners of uttering under specific conditions of communication. Speech type varies along multiple dimensions, including the number of and relationship between conversational participants, the degree of intelligibility required [3], the degree of preparedness of the discourse, etc. Recent research [3], [4], [5], [6] has demonstrated how some prosodic characteristics regularly vary according to specific dimensions of the situation (for example, F0 register span increases in public communication).

Once prosodic profiles are established for each targeted speaking style, they are compared with the prosodic profile of the existing neutral voice produced by the TTS system. One-to-one differences in prosodic parameters are then used to modify the synthetic voice, using several procedures, either within the TTS system or during a post-processing phase.

2. Text-to-Speech system eLite

The eLite TTS system operates by selecting non-uniform units from a speech database of 56,000 diphones. The selection algorithm is designed so that it minimizes the cost of the target (based on linguistic characteristics) and the cost of the join (based on acoustic characteristics). The original prosody of the units is preserved, resulting in a quite natural, if somewhat monotonous and neutral melody and rhythm (see [7], [8] [9]). Consequently, the resulting prosody depends strongly on the original voice and style in the database (in this case, an unemotional reading style).

Any modification of the synthesized prosody is therefore made difficult because of the architecture of the system. In other words, "the quality of the synthesis relies on the fact that little or no signal processing is done on the selected units, thus the style of the recording is maintained in the quality of the synthesis. The synthesized style is implicitly the style of the database." [10]

3. Analyzing speaking styles

What makes a speaking style perceptually different from another one, in a salient way? Various studies [3], [4] have shown the prosodic parameters on which speaking styles typically and regularly differ.

We used the ProsoReport tool [11] to extract quantified information about each style sample. This tool requires the preliminary processing of data using the following Praat-run scripts [12]: the EasyAlign tool segments a recording into phones, syllables and words on the basis of speech signal and orthographic transcription; Prosogram [13] is used to segment and stylize F0 into perceptual nucleic tones based on syllable segmentation; the ProsoProm tool [14] automatically determines which syllables are prominent; eventually the ProsoReport tool gathers prosodic measurements into a table containing a list of ca. 70 prosodic descriptors.

We applied the complete description procedure to the output of the TTS system (representing 20 minutes of synthesized speech) and to 9 recordings (3 speech samples for each of the 3 speaking styles, representing 10 minutes per style) and established a list of the most salient differences between each style and the neutral synthesis produced by eLite.

As for rhythm (Table 1), the following features turned out to be relevant for distinguishing speaking styles:

- speech rate (number of syllables uttered per second, pauses excluded) is fast in Radio News, intermediate in Conversation and quite slow in Political Address;
- the pause rate greatly varies between Political Address (31.67% of the speaking time is pausing) and the two other styles;
- the mean number of syllables between two pauses varies from 8 (in Political Address) to 15 or 16 in Radio News or Conversation and is responsible for Interpausal Units [15] of highly diverging lengths;
- within the silent pauses, we calculated the number of pauses with audible breath, in order to insert breaths within the synthesized voice, which is hardly ever the case in TTS synthesis;
- the rate of schwa deletion in final-word position is higher in informal, conversation style (80%) than in public style (around 57%);
- the proportion of hesitation particles (like "euh" in French) ranges from 0.05% in Political Address to 7.51% in Conversation.

Table 1. Differences between speaking styles: **Rhythm**
(TTS, News: Radio News, Pol: Political Address, Conv: Conversation)

	TTS	News	Pol	Conv
Speech Rate (syl/sec)	**5.6**	5.8	4.8	5.3
Pausing Time (%)	**26**	10 97	31.67	16.73
Mean Nb. of Syllables between Pauses	**8**	15	8	16
Pauses with Breath (%)	**0**	58.5	33.4	57.2
Rate of Final Schwa Deletion	**60.3**	57	57.35	80
Hesitation Syllables (%)	**0**	1.83	0.05	7.51

One step further in the analysis of speaking style, we identified prominent syllables - that is, possibly accented syllables - by means of a speaker-independent automatic detection procedure [16], [17] that considers relative height and duration of syllables. Syllables detected as prominent stand out against their local environment because of an extra-long duration, a higher F0 mean or a rising pitch movement within the syllable. Measurements reveal that (see Table 2):

- The neutral synthesis has the lowest rate of prominent, accented syllables (15.5%), while Political Address has the highest rate (27.5%).

When combined with grammatical annotation [18], prominence detection results in a categorization of final accents (on the last syllable of a full lexeme, namely a noun, an adjective, a verb or an adverb) and initial accents. Final accents contribute to segment the flow of speech into prosodic units, whereas initial accents create emphasis (the so-called 'didactic style' or 'insistence accent' [16], [19]).

- Political Address has both the highest rate of Final and Initial Accents, resulting in a rather emphatic style, with short prosodic units (it also has the smaller mean number of syllables within two pauses, which amounts to 8); Radio News style has longer prosodic units with as many Initial Accents as in Political style, while Conversation has the longest prosodic units and the fewest initial accents.

Finally, the melodic register was measured for each style, using semi-tones (instead of Hz), which makes it possible to compare registers between speakers, regardless of whether they are male or female. The most monotonous voice is the synthetic voice, with only 4.3 semi-tones between the lowest and the highest pitch values (excluding the 5%-95% extreme parts of the register). Political Address and Radio News both make use of a wider register, which has been demonstrated to be typical of broadcast discourse [4] but [5].

Table 2. Differences between speaking styles:
Intonation and Accentuation

	TTS	News	Pol	Conv
Prominent Syllables (%)	15.5	25.4	27.5	20.4
Final Accents (%)	34.01	42.75	48.97	32.39
Initial Accents (%)	5.13	22.60	22.88	14.86
F0 Range (in ST)	4.3	10.5	10.5	7.4

In the next section, we explain how we modified the TTS eLite system in order to accommodate stylistic variation.

4. Implementing various speaking styles within the eLite TTS system

Speech generated by selecting and joining non uniform units is likely to be more natural (see Section 2). However, the prosodic analysis carried out in Section 3 shows that neutral synthesis voice is monotonous, less expressive and has less contrast between accented and non-accented syllables.

The gap to be filled between neutral synthesis and political, broadcast news or conversational speaking style can be calculated from Table 1 and Table 2, yet not every modification in the speech signal can be carried out in post-processing without deteriorating the signal.

Some of the modifications in rhythm (reported in Table 1) were implemented in the TTS itself. Modifications in accentuation and F0 range were implemented in a post-processing treatment.

4.1. Within TTS processing

Prior to selecting units from the database with the purpose of generating the speech signal, the following operations are required: prediction of the insertion of pauses, breaths and hesitations; prediction of schwa deletion.

- **Pauses**: the pauses of each speaking style were analyzed (mean duration and standard deviation) and showed distributions with 2 modes, whose durations were respectively attributed to short and long pauses in the TTS system (see Table 3 and Figure 1).

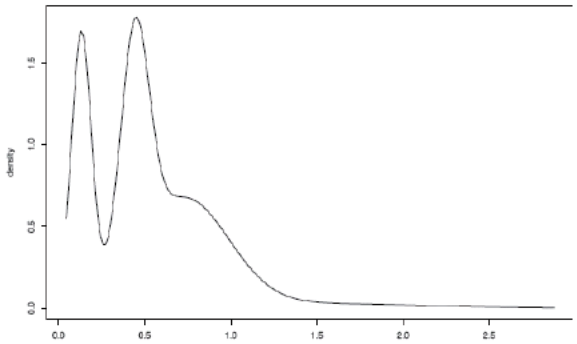

Figure 1. Mixture model of distribution of pauses for conversational speaking style.

Table 3. Mean Duration and Standard Deviation of Pauses within the Three Styles

	TTS	News	Pol	Conv
Mean Duration of Short Pauses (Std Deviation)	500 (0)	96 (1)	426 (24)	287 (4.5)
Mean Duration of Long Pauses (Std Deviation)	1100 (0)	404 (15)	1050 (215)	728 (66)
Mean Duration of Breaths	--	430 (190)	430 (190)	650 (340)

- **Breathing**: some breathing samples from the authentic speech recordings in the same speaking style were extracted and added to the database. The reason for doing this is that no breathing sound could be found in the TTS database and that the intensity, spectral properties and duration of the breaths within the natural speech corpus were specific to the speaking style. Breathing sounds were inserted in the synthetic speech according to the results from a previous study on frequency and duration of breathing for each speaking style [18].
- **Noise:** light white noise was added in order to fill synthetic silent pauses.
- **Hesitations**: only 2 samples of French "euh" ("er") were found in the original recordings of the TTS database. Hesitations are inserted according to the frequency found in the authentic corpus. As it turns out, only conversational style had hesitations (1 out of 13 syllables).

Those modifications result in a synthetic voice whose characteristics are closer to original styles. Modification in speech rate, F0 and accentuation was done in the post-processing step.

4.2. Post-processing

Basically, speech rate and F0 register may be modified at a global level, which means that one can uniformly reduce the duration of syllables (for speeding up the speech rate) or extend the distance between the extreme F0 values, with the purpose of increasing the "melodicity" of the voice.

An in-depth analysis of the speech material including our 3 original styles – journalistic, political and conversational – revealed that syllables within a given style behave differently according to their location within grammatical units and intonation units. For example, when the syllable is located at a prosodic boundary, its duration may be emphasized by 163 to 199% depending on the speaking style (see also [6]).

On the other hand, our prosodic analysis showed that speaking styles differ in the proportion of syllables with a final or initial accent. Post-processing was then conceived for generating a satisfying rate of initial and final accents (syllabic prominences) and for modifying the mean duration and mean F0 of accented and unaccented syllables, so that they match the prosodic characteristics of the original style. For this, we used the well-known overlap-and-add technique implemented in Praat [12].

4.2.1. Modifying 6 categories of syllables

Each syllable belongs to one of the following three types: word-initial, word-median and word-final. Syllables from clitic words (like determiners or weak pronouns) are handled like median syllables. Prominence detection has the added benefit of categorizing each syllable as prominent or not. When combining syllable position information with prominence detection, we were able to distinguish 6 categories of syllables. For each category, a mean relative F0 and duration were computed. Thus, 12 coefficients describing the 6 syllable types were found for each style, as well as for the neutral synthetic voice.

The main idea of the style conversion system is to apply to each of the 6 syllable categories the difference between the mean relative F0 of a style to imitate (in semitones) and the mean relative F0 of the neutral synthetic voice. The same applies, but as a ratio rather than a difference, to the relative duration. Every F0 or duration modification was local (since it applied to a syllable according to its location), but resulted in a global modification of speech rate and F0 register (see Table 5).

4.2.2. Modifying the rate of final and initial accents

The other central parameter typical to each style is the proportion of initial and final accents. Adding accented syllables to the neutral synthetic voice was done by creating an additional category of syllable: syllables can be prominent, non-prominent, or "to-be rendered prominent" . The latter category is acoustically modified by using the F0 and duration parameters of a prominent syllable in the targeted style, as compared to the synthetic neutral non-prominent syllable. The syllable to which this modification is applied is chosen according to the difference of prominence rate in syllable position (initial or final). For instance, 32% of the final syllables in the synthetic voice were found to be prominent, whereas 43% of the final syllables turned out to be prominent in the Radio News style, as shown in Table 2. Thus we had to make prominent 1 out of 6 non-prominent syllables [as (1-0.32)/(0.43-0.32)] of the neutral TTS speech to raise the prominence rate to that of the Radio News style. The same applies to the initial accented syllables ratio.

Table 4 summarizes in grey the percentage of prominence in initial and final position for the 3 speaking styles and for the neutral TTS. The number following an arrow (→) indicates the syllabic rate at which a prominent syllable has to be added.

Table 4. Percentage of prominent syllables at initial (I) and final (F) word position (in grey columns) and rate of prominent syllables to be added (→) during the conversion of the synthetic voice into a specific speaking style

	TTS	News	TTS to News	Pol	TTS to Pol	Conv	TTS to Conv
I	**5**	23	→5	23	→5	15	→9
F	**32**	43	→6	49	→4	34	→0

Table 5. Coefficients for the conversational style. Note that the rate of "to-be-prominent" final syllables is 0 as the percentage of final accents is lower in original Conv style than in neutral TTS speech

		Rate	F0 (ST)	Duration
I	*non-prom*		-0.58	0.99
	prom		-0.42	1.17
	to-be prom	9	2.55	1.64
F	*non-prom*		0.02	0.91
	Prom		1.57	0.88
	to-be prom	0	2.12	1.7
M	*non-prom*		-0.24	1.05
	Prom		0.14	1.31

All in all, our conversion system is based on the calculation of 18 coefficients: 12 coefficients for F0 and duration of prominent/non-prominent X initial/median/final syllables, 2 rates of non-prominent syllables that are to become prominent; 4 coefficients for F0 and duration for these "to-be-prominent" initial and final syllables. Table 5 summarizes the set of coefficients for obtaining the conversational speaking style from neutral synthesis.

5. Experimental validation

Validation was carried out by verifying that some prosodic characteristics of the synthesized "stylized" speech were modified toward the prosodic characteristics of the targeted original style. This closed circuit validation is acceptable, as we evaluate local modifications with global measures. More precisely, the synthesized speech in the three styles was analyzed as the natural speech was through the steps described in §3. Table 6 shows the parameters that could be computed this way. The speech rate (in syl/sec) was increased from 5.6 to 5.7 for the News style, and lowered to 5.3 and 5.2 for the Pol and Conv style (without reaching the targeted speech rate, which was at 4.8 and 5.2 respectively).

Table 6. Validation by comparing prosodic features in the original (upper lines) vs. synthetic styles (below).

	TTS	News	Pol	Conv
		TTSNews	TTSPol	TTSConv
Speech Rate (syl/sec)	5.6	5.8	4.8	5.3
		5.7	5.3	5.2
Pausing time (%)	26	11	31.7	16.7
		9.1	29.7	17.3
Nb. of Syllables between Pauses	8	15	8	13
		10	8	8.5
F0 Range (in ST)	4.3	10.5	10.5	7.4
		7.1	8	5.4

The same applies to the pausing ratio, the number of syllables by interpausal segments and the F0 range. As final remarks, we can say that combining the two techniques (within the TTS and post-processing adjustments) does not make it possible to completely reach the targeted parameters. At first listening, the signal modifications do not degrade naturalness and seem to add expressivity. This has to be confirmed with a perceptual validation.

6. Discussion and conclusions

As far as the originality of our approach is concerned, two issues in particular deserve mentioning here. On the one hand, we added "naturalness" to the synthetic voice by adding some breathing to silent pauses, better modeling the length of silent pauses and adding hesitation particles (for conversational style). On the other hand, we modified speech rate or F0 range not on a global basis (by uniformly modifying the duration or F0 of each syllable) but locally, by modifying each syllable type so that it would be closer to the same syllable type (initial accent, final accent, unaccented syllable, etc.) in the targeted style. The local changes altogether not only made specific changes on 6 types of syllables but also made the global parameters come closer to those of the targeted speech.

The results of the present analysis being as encouraging as they are, future research will focus on the following aspects: addition of new styles, annotation of syllabic prominence in the database for enhancing the selection of units according to their accentuation, better technical integration of the style converter within the TTS system and perceptual validation.

7. Acknowledgements

This research was supported by grant no. 0616422 from the Walloon Region of Belgium for project: "Expressive. Système automatique de diffusion vocale d'information dédicacée: synthèse de la parole expressive à partir de textes balisés".

8. References

[1] Beaufort, R. and Ruelle, A. "eLite: système de synthèse de parole à orientation linguistique". Proc. of JEP, 509-512, 2006.

[2] Hirschberg, J. "A Corpus-based Approach to the Study of Speaking Style", in M. Horne [Ed]. Prosody: Theory and Experiment. The Netherland, Kluwer, 335-350, 2000.

[3] Eskenazi, M., "Trends in Speaking Styles Research". Proceedings Eurospeech, Berlin, 501-509, 1993.

[4] Simon, A.C., Auchlin, A., Avanzi, M. and Goldman, J.-Ph. "Les phonostyles: une description prosodique des styles de parole en français", in M. Abecassis and G. Ledegen [Eds], Les voix des Français. En parlant, en écrivant. Berne: Peter Lang, under press (2009).

[5] Llisterri, J. "Speaking styles in speech research", ESLNET/ESCA/SALT Workshop on Integrating Speech and Natural Language, Dublin, 1992. [available at http://liceu.uab.es/~joaquim/publications/SpeakingStyles92.pdf]

[6] Abe, M. and Mizuno, H. "Speaking styles conversion by changing prosodic parameters and formant frequencies", *ICSLP 94*, Yokohama, Japan, 1455-1458, 1994.

[7] Colotte V., Beaufort R. "Linguistic Features Weighting for a Text-to-Speech System Without Prosody Model", Proceedings of Interspeech 2005, 2549-2552, 2005.

[8] Colotte V., Beaufort R., "Synthèse vocale par sélection linguistiquement orientée d'unités non-uniformes: LiONS", Proceedings of JEP'04, Fès, Maroc, 4 p., 2004.

[9] Latacz, L., Y. Kong, and Verhelst, W. "Unit selection synthesis using long non-uniform units and phoneme identity matching," Proceedings of the 6th ISCA Workshop on Speech Synthesis, 270–275, Bonn, Germany, 2007.

[10] Black A. "Unit selection and emotional speech", Proceedings Eurospeech, Genève, 1649–1652, 2003.

[11] Goldman, J.-Ph., Auchlin, A., Simon, A.C. & Avanzi, M. "Phonostylographe : un outil de description prosodique. Comparaison du style radiophonique et lu". Nouveaux cahiers de linguistique française 28, 219-237, 2007.

[12] Boersma, P. and Weenink, D. "Praat: doing phonetics by computer (Version 5.1.20)" [Computer program]. Retrieved October 31, 2009, from http://www.praat.org/

[13] Mertens, P. "The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model". In B. Bel & I. Marlien [Eds], *Proc. of Speech Prosody*, Nara, Japan, 2004.

[14] Goldman, J.-Ph., Avanzi, M., Simon, A.C., Lacheret, A. and Auchlin, A. "A methodology for the automatic detection of perceived prominent syllables in spoken French". Proceedings of Interspeech 2007, 98-101, 2007.

[15] Goldman, J.-Ph., Auchlin, A. and Simon, A.C. "Description prosodique semi-automatique et discrimination de styles de parole". (submitted)

[16] Lacheret-Dujour, A. and Beaugendre, F. La prosodie du français. Paris, CNRS, 1999.

[17] Avanzi, M., Goldman, J.-P. Lacheret-Dujour, A. Simon, A.-C. & A. Auchlin, "Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé", Cahiers of French Language Studies, 13/2, 2-30, 2007

[18] Roekhaut, S. "Expressive. Système automatique de diffusion vocale d'information dédicacée: synthèse de la parole expressive à partir de textes balisés", Scientific Report, Unpublished ms, September 2009.

[19] Léon, P. Précis de phonostylistique. Parole et expressivité. Paris, Nathan, 2003.