The Continuous and Discontinuous Styles in Czech TTS

Tomáš Duběda[1] and Jiří Hanika[2]

[1] Charles University, Prague, Faculty of Arts, Nám. J. Palacha 2, 116 38 Praha 1
dubeda@ff.cuni.cz
[2] Charles University, Prague, Faculty of Mathematics and Physics,
Malostranské nám. 25, 118 00 Praha 1
geo@cuni.cz
http://epos.ure.cas.cz/

Abstract. In this paper, a partial prosodic analysis of two sets of speech styles in Czech is carried out. Four styles have been chosen as representatives of discontinuous styles (Address, Spelling, Headings, Dates and Prices); three styles have represented continuous styles (Weather Forecast, Sport News, News). Based on this analysis, the prosody of the selected speech styles has been simulated using the Epos[3] speech system.

1 Introduction

Presently, the text-to-speech conversion as such is already a well-solved problem. Working speech synthesizers are available for hundreds of languages and many complete TTS systems have reached the commercial applicability stage, at least for major languages. Only few of them however produce a natural output, partly because the subjective requirements for naturalness in the speech community are less benevolent than they were ten years ago. Worse still, many applications of TTS are from a grammatic point of view very eccentric and there are little or no detailed descriptions of human speaker behavior under such conditions. Given a text in a Latin-type writing system, the open problems reside almost exclusively in the prosody domain.

It is well known that not every aspect of usual human behavior is suitable for simulation in TTS; for example, the Czech word stress, which is partly intensity-based, has been found very tricky to implement automatically in the same way, because some occurrences are frequently misinterpreted as a semantic emphasis or even emotional speech by listeners. On the other hand, a pure pitch-based approach led to correct and relatively natural results. That's why building a description for a given speech style is an iterative process; if too close an imitation fails, reasons and alternative ways have to be looked for. Our research has been verified using the Epos speech system[3], which has been originally designed exactly for the task of testing hypotheses in prosody modelling in TTS, and whose standard prosody has been configured to use the prosody modelling strategy described in [5].

2 The Styles of Speech

We have chosen the term "speech styles" to designate different types of situations in which speech synthesis is used, as well as the associated speech behavior of human speakers. Thus, the term does not correspond to a classical, linguistic definition: in this article, the speech styles are characterized by the type of text, its syntax, conveyed information, interactivity and similar factors. As most factors contributing to a speech style are completely text-based, a TTS system has to focus on the prosody modelling to perform well in a given speech style.

In many cases, a non-specific speech synthesizer is primarily designed for continuous, fully grammatical speech (sequence of sentences with their transphrastic relations), e. g. a news article. However, to meet with the demands of everyday life, many applications have to cope with tasks that differ from such non-specific text, the task being a sequence of phone numbers, short answers and alerts (Yes, No, Continue, Repeat) or anything else. To study the prosody of these language productions, we used a corpus of natural speech recordings made by the participants of the COST 258 project. The corpus includes 13 languages spoken by 17 speakers. For each language, a set of 15 short texts had been recorded, ranging from a proper name spelling to an interactive database query. For Czech, one male speaker (a rather grave voice) was recorded.

Our aim was to detect relevant prosodic cues that mark the chosen speech styles different, to quantify them, and to incorporate their models into the non-specific prosody module as described above. Our starting point was necessarily the above mentioned set of prosodic rules for Czech, and we also tried to use standard facilities of this rule set in all cases that option seemed adequate. The modelling was based on data that came from the analysis of a single speaker, which means that it emulates *his* speech behavior rather than "average" Czech speech behavior. We consider this approach to be well-founded, since the speaker's prosody is fully acceptable. The phonetic analysis was performed using the SpeechStudio software by Petr Horák.

3 Continuous and discontinuous styles

The language (performance) can be seen as a continuum of texts ranging from the most central (i.e. fully grammatical) spoken texts to the periphery of language, and so do speech styles. The core of a natural language is associated with prosodic features which have been often subjected to a phonetic analysis; on the other hand, many if not most speech-enabled systems focus on a very different application domain: dialogue systems with limited or formalized output, screen readers for blind users, or speech interfaces of monitoring systems. The speech styles employed by human speakers in such circumstances are often termed discontinuous, because the prosodic utterance typical for such speech style is shorter than a sentence, and from the grammatical point of view is thus incomplete. On the other hand, the more classical speech styles differ only slightly in the repertoire of prosodic devices applied, because the structural differences of texts found there are minor as well. In fact, a "standard"[5] prosodic analysis for the language in question suits continuous styles well, and that's why a standard text-to-speech processor handles these styles acceptably. Discontinuous styles show three important features:

1. Their application domain tends to be specialized and narrow, which makes them easily formalized, while less linear,

2. As the syntactic rules are relaxed within a discontinuous style, the phrases are very short and redundancy is suppressed; that is particularly unfortunate, as numerals, proper names and other sensitive texts are likely to occur in some of these styles.

3. Staying at the periphery of language, they differ from each other considerably.

For these reasons, the analysis of a continuous speech style is a very different task from that of a discontinuous speech style.

4 Discontinuous styles

A superficial auditive comparison of the recorded speech sequence and its synthetic imitation in the automatic TTS synthesis designed primarily for continuous texts was used to determine the prosodic features upon which the model should be based: terminal/non-terminal F_0 contours, insertion of pauses, F_0 declination and timing. Some other proposed features were rejected, as their significance in the recorded speech sequence was unclear. Four styles in this category have been studied: "Address", "Headings", "Dates and Prices" and "Spelling".

Some of the styles in question have a rigid structure (e. g. a name – address – phone number sequence generated by a database), but show no 'real' syntax and mostly lack any punctuation at the input. It is thus necessary to insert additional punctuation marks into the synthesized speech sequences in order to obtain a suitable prosody reflecting the 'pseudo-syntax' of the utterance.

The insertion of a comma implies a non-terminal contour, the insertion of a stop implies a terminal contour. According to our auditive evaluation, the insertion of punctuation marks is sufficient to generate satisfactory intonation contours. The automatic duration increase and intonation decrease at the end of prosodic boundaries (which is a quasi universal prosodic feature in Czech) seems to be sufficient too.

The insertion of commas and stops also implies the presence of pauses. For numerals, a clear and slow realization is required, because the information is very compact and there is no redundancy in the utterance.

Thus, the sequence first name – last name – street – block – zip code – city – phone number in the "address" style was split with additional punctuation marks that imply the corresponding pause and intonation contour generation rules as follows:

(database output)
```
Josef Novák
Národní třída 16
110 00 Praha 1
21 63 89 15
```
(prosody modelling; other TTS transformations not shown)
```
Josef Novák, [pause 480 ms/rising contour] Národní třída 16,
[pause 480 ms/rising contour] 110 00, [pause 480 ms/rising
contour] Praha 1. [pause 640 ms/falling contour] 21,
[p. 480 ms/rising] 63, [p. 480 ms/rising] 89,
[p. 480 ms/rising] 15. [p. 480 ms/falling]
```
The "spelling" style requires special treatment for F_0 declination (gradual pitch descent through the sentence). Unlike the declination used with continuous speech, the spelling declination is much more abrupt.

5 The continuous styles

Three representative styles with a high frequency have been chosen out of the spectrum of continuous speech styles defined by the COST 258 project: non-specific news report (N), weather forecast (WF) and sport news report (SN). It has been said above that continuous styles, staying in the centre of the language system, show less variation and can be generated by the non-specific prosodic module more satisfactorily than discontinuous styles. In fact, the observed differences are often very subtle, and the degree of acceptability for the three styles is rather high, even if only the non-specific prosody modelling is employed.

We focused on the following properties:

- Speech and articulation rate
- Length of prosodic units (stress unit and intonation unit)
- Pauses
- Timing of final clauses
- F_0 contours in final position

The units of prosodic segmentation that our synthesis operates with are stress unit (SU, a chain of syllables with an initial stress, mostly comprising of at most one semantic word and zero or more grammatical words), and intonational unit (IU, a chain of stress units terminated by a phonological F_0 contour, corresponding mostly to an inter-punctuation stretch or clause). Let us note at this place, that our hierarchy of linguistic units such as clause or sentence is also methodologically based on the TTS application, i.e. the usual punctuation and transformational rules are used to delimit such units. This attitude is a special case of using the standard facilities of our base rule set where possible; for our texts however — and for almost all of Czech in general — the "graphical" units coincide with their prosody description counterparts, i.e. there happens to be no discrepancy between the two points of view.

The average articulation rate (syllable count divided by articulation time, i. e. without pauses) is about 5.23 syll/s and doesn't depend on the speech style chosen significantly. (This is however not always true locally; special attention must be paid to clause-final stress units.)

The length of sentences and intonation units (IU, inter-punctuation stretch of stress units) is as follows:

Table 1. Lengths of prosodic units

Text	N	WF	SN
Average sentence length (syll.)	61,6	27,3	19,9
Average IU length (syll.)	23,2	11,3	12,2

The N text shows very long sentences and intonation units; the other two texts comprise of more clauses, especially the SN text, which is the most dynamic and expressive one. This observation has no direct effect on our TTS prosody modelling, while it gives some insight into the differences observed among the texts. It is also possible to employ the style difference — if any — in treating certain punctuation-less constructions as clause terminators in some cases, for example, but such heuristics are hard to propose, test and evaluate.

The average length of a stress unit is about 2.73 syllables for all the three styles, i.e. the basic characteristics of the texts are comparable.

Table 2. Pauses

Text	N	WF	SN
Total number of pauses	10	17	16
Frequency of pauses (syllables/pause)	18.5	11.3	9.9
Average pause duration (ms)	552.9	684.1	522.9
No-punctuation pause average duration (ms)	146,0	88,0	425,0
Average duration of a pause after a comma	357,0	357,0	235,0
Average duration of a pause after a stop	1652,0	1329,0	958,0

The frequency of pauses increases in the order SN-WF-N and it is inversely proportional to the length of a suprasegmental unit for physiological reasons. Spontaneous pauses (pauses corresponding to no punctuation mark or any other clause terminator, which leaves 16 be ignored at the present stage, because no complex syntactic analysis is performed by Epos. Therefore, pauses can be safely inserted almost exclusively at clause-marking or sentence-marking punctuation. The values in the "Average duration of a pause after a comma" and "Average duration of a pause after a stop" rows were directly inserted into the corresponding prosodic rules.

The timing was studied only in terms of clause-final lengthening:

Table 3. Clause-final lengthening

Text	N	WF	SN
Average syllable duration of a clause-final SU (ms)	196	253	223
Average syllable duration of a sentence-final SU (ms)	186	179	182
Clause-final / sentence-final SU syllable duration ratio	1.05	1.41	1.23

The table shows that the clause-final lengthening increases in the WF-SN-N order.

Observe that sentence-final stress units are actually shorter than clause-final ones in general. This unexpected phenomenon is indicative of information-dense speech styles such as weather forecasts, where phrasing and speech are both particularly important. There, a pause indicates a sentence-level break, while an overall slowdown, *including the pause*, marks a clause-level break. That way, lengthening proper is not a phrasing marker in presence of a pause and can safely be somewhat suppressed. This effect is obviously not critically dependent on average sentence length. As far, as the SN style is concerned, a limitation of our corpus may come into play: a real-life SN production may be much less planned and syntactically coherent, and pausing and phrasing is likely to behave differently.

Modelling of such style-dependent differences is straightforward.

References

1. Dohalská, M., Duběda, T.: Experiments with Duration and Intensity Variation in Czech Text-to-Speech. Proceedings of the 6th Czech-German Workshop on Speech Processing, Prague, September 1996
2. Dohalská, M., Duběda, T., Mejvaldová, J.: Speech Styles in Synthetic Czech. oral presentation, COST 258 Executive Committee meeting, Budapest, September 1999
3. Hanika, J., Horák, P.: Epos - A New Approach to the Speech Synthesis. Proceedings of the First Workshop on Text, Speech, Dialogue — TSD'98, Brno, Czech Republic, September 1998, pp. 51–54
4. Hanika, J., Horák, P.: Dependences and Independences of Text-to-Speech Hans-Walter Wodarz (ed.): Forum Phoneticum 69, Frankfurt a.M., to appear
5. Palková, Z., Ptáček, M.: Prosodic Issues in Czech: An Application in TTS. Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm 1995, vol.4, pp.380-383