PROSODIC ISSUES IN CZECH: AN APPLICATION IN TTS

Zdena Palková and Miroslav Ptáček
Institute of Phonetics, Prague, Czech Republic

ABSTRACT

The Czech prosody algorithm in TTS diphone synthesis is based on segmentation into stress units. It is sensitive to their positions in sentences; it creates separate sound patterns for duration, and for the F0 contour. Results confirm a hypothesis that prosodic modulation of a Czech text is well conceivable in terms of a string of linear units having characteristic sound contour, rather than as a recurrence of prominences on a neutral backdrop.

1. BACKGROUND

The purpose of research is to verify hypotheses concerning prosodic properties of Czech. Designing a plausible algorithm for the suprasegment level of the text-to-speech diphone synthesis is a means to this end, and has also a practical goal.

The description of Czech prosody can get a suitable footing in a three-tier hierarchy model of linear suprasegmental units: *syllable - stress unit* (a group of syllables having a single word stress) - *intonation unit* (a group of stress units joined by intonation). Some properties of Czech make this conceptual framework applicable at explaining general phonological issues in the prosody of Czech, and also at analysing individual text utterances, [1].

The approach adopted in our present research is also underpinned by a hypothesis that assumes as primary the very fact of segmenting the string of syllables, i.e. grouping syllables to stress units and stress units to intonation units. The search is for sound means that are instrumental in such grouping, and also for perceptual boundary signals. In other words, the linear rhythmic units do not a priori derive from prominences (word or sentential accents) as in traditional approaches. This approach responds to earlier experiments in search of the acoustic features making the nature of the word stress in Czech, [2].

(Concerning the Czech word stress, basic structural descriptions of Czech characterise it as fixed on the first syllable of words while the difference of stressed and unstressed syllables has no impact on quality or quantity of vowels.)

The key unit of the present description is the stress unit, which makes the intermediary tier, as it can reflect projections of some syllabic features, and also some features of its intonation unit. Furthermore, the unit is rather easy to mark out in Czech written texts since it has close links to the word. The necessity of getting support from phenomena definable in written texts by formal analysis follows from features of a text-to-speech production, and some trade-offs have to be done because of it, [3].

2. PROCEDURE

The following principles apply in the TTS algorithm.

a) The stress unit position is related to the text structure. Ideally, the relevant superordinate unit would be the intonation unit. Its determination in a particular text depends, however, also on syntactic and semantic information contingent on a larger context. For the purpose of automated synthesis, a "syntactic unit" is being used, which is defined as the text enclosed within two punctuation marks, in terms of Czech orthography (slightly adjusted), as the punctuation in Czech conforms to formal rules motivated largely by syntactic sentence relations.

The following additional distinctions are made within those discriminated units if necessary:

(I) - the initial position, the first stress unit in a syntactic unit

(M) - the medial position, not bordering on a following syntactic unit boundary; another stress group follows that belongs to the same superordinate unit

(F) - the position preceding the boundary signal of the syntactic unit that is not the last one in the utterance (i.e. the position preceding a comma, or a colon)

(FF) - the position preceding the boundary signal of the syntactic unit that is the last one in a finished utterance (i.e. preceding such marks as . ; ? !).

The FF-F-I-M order of preference is used when evaluating positions in the bordering versions of short texts.

b) Acoustic features are set separately for stress units of various lengths.

c) Acoustic features of stress units consist of a number of acoustic qualities, and their production algorithms have also separate designs.

No special features have been set out for initial syllables as "stress" bearers. The pattern of acoustic features spreads across the whole of stress unit. Duration modification of segments (phones or diphones) in stress units, and modification of the F0 contour, are fundamental. Modification of the dynamics is reserved for phenomena from higher levels of the text structure (emphasis, etc.); so far it has not been worked out in our automated synthesis.

Provisions have been made to allow for single acoustic features to modify the chain of stress units in various ways related to their positions within the sentence unit. E.g. a one-syllable word in the I position combines with the following word into a single unit through modifying duration but not through F0.

3. SEGMENTATION INTO STRESS UNITS

A text containing sequences of polysyllabic words submits in Czech successfully to the hypothesis that each word makes a stress unit. Variability is introduced by a one-syllable word which can stand apart, or tie to the neighbouring words. In natural speech, the solution depends on semantic and pragmatic aspects.

An algorithm has been designed to enable for such monosyllables to get connected in larger stress units based solely on the unit length and position within the syntactic unit. Tendencies found in spontaneous natural speech are made use of: The length of the produced units is 6 syllables or less, and when segmenting longer chains of syllables, the parting is either symmetrical or leaves the first portion longer. Special rules apply for single monosyllables in the I and F positions. Some exceptions have to be stored. Infrequent cases of a one-syllable word in a position "unstressed" by rule when expected to bear emphasis from its context still present a challenge.

4. DURATION PATTERNS

Duration of phones within stress units is probably influenced by the number of syllables and their types (presence or absence of a coda). Syllable breaks in Czech, however, are difficult to detect due to an abundance of consonant clusters. That is why differences in segment duration are derived in synthesis simply out of the number of phones in the stress unit.

Listening tests showed that, in perception of stress units in continuous speech, acceptable alterations in average duration of phones reflect the relationship

$$T(n) / T(m) = (m/n)^{0.12}$$

where $T(n)$ is the average phone duration in a stress unit containing n phones, and $T(m)$ is the average phone duration in a reference stress unit containing m phones. The reference number of phones has been set to 5, i.e. $m = 5$, as the stock of diphones was extracted mainly out of 5-phone words. If a relative value of 100% is set to the average phone duration in a 5-phone unit, the average phone durations get values as in tab. 1. *(See the table next page.)*

In accordance with results of listening tests, the phone durations in stress units longer than 12 phones have been exposed to no further reductions. Experiments have also shown that parameters of duration differences found out in isolated words cannot be applied. Such words reach their exponent value of 0.41 while keeping an analogous relationship between duration and a number of phones in the stress unit, [4]. The fact that larger differences in duration of phones are not acceptable in the

Tab. 1. Duration of phones related to their number in the stress unit.

Number of phones in a stress unit:											
1	2	3	4	5	6	7	8	9	10	11	12
Relative duration of a phone:											
120	111	106	102	100	98	96	94	93	92	91	90

perception of continuous speech is in agreement with the usual ranking of Czech with the syllable-timed languages.

5. PITCH PATTERNS

The F0 contour in stress units is made up separately for I+M+F and FF positions.

5.1 Stress Units in the I+M+F Positions

a) Selection of fundamental pitch patterns (i.e. F0 contours) has been based on earlier experiments which required for listeners to break up continuous flow of sound into stress units lacking any possible semantic clues [5]. Particular configurations of pitch patterns within types have been established through selection from a larger number of variants with regard to acceptability for listeners.

b) Separate pitch patterns have been set up for each stress unit of a particular number of syllables. Each type distinguishes two sets of variants (containing 2 - 6 patterns each):

(A) an off-falling contour (the last change is the fall, the pitch pattern stops at the same or lower tone related to its first syllable),

(B) an off-rising contour (the last change is the rise, the pitch pattern stops at the same or higher tone related to its first syllable).

A level contour (i.e. lacking any change of pitch) is used only with two-syllable stress units, and operates as a member of the A set.

A maximum change of pitch within a pitch pattern is 5%.

c) Rules have been accepted for chaining pitch patterns along stress unit sequences in specific texts. They consist mainly of ongoing alternations of variants from the A and B sets while the selection of a particular pitch pattern is unconstrained.

d) The phonologically relevant F position is implemented with a distinctively linked pitch pattern, usually through lowering the initial syllable of the stress unit by -2% against the closing syllable of the preceding unit.

e) Auxiliary rules have been established to solve some specific situations, e.g. for monosyllables in the F and I positions.

f) Possible excessive drop or soar of F0 in a longer chain of stress units has so far been dealt with by means of follow-up corrective rules applicable at the end of each stress unit in the FF position, i.e. after modulation in the concluded utterance have been created.

5.2 Stress Units in the Closing FF Position

a) Again, separate pitch patterns have been set up for each stress unit of a particular number of syllables. Changes in pitch have been verified by listening, and do not exceed 15%. The A set of pitch patterns (2 - 4 patterns of each length) are in force preceding the punctuation marks . ; ! and some of the ?. The B set of patterns (2 patterns of each length) apply before ?. A pitch pattern selection in questions is necessary because the F0 contour is relevant in yes-no questions in Czech. A decision has to be based on a set of rules detecting key words stored in advance.

b) The link of a stress unit in an FF position is controled by a rule sensitive to the preceding stress unit F0 contour.

6. CONCLUSIONS

The algorithm that has been worked out provides our synthesis of Czech with a prosody in a well acceptable shape. Infrequent mis-modulations arise from the inability to apply wider context-based semantic information via formal rules. Occasional corrective signals to non-standard segmentation or pitch pattern placement have to be introduced so far on an ad hoc basis.

The results seem to support a hypothesis that prosodic modulation of a stream of syllables in Czech is well conceivable in terms of a string of linear units having characteristic sound contour, rather than as a recurrence of a prominent syllable on a neutral backdrop.

Further research is now going to focus on the issue of analogous relationships within the higher intonation unit, and on reaching their formal description. It appears possible to investigate implications of such an approach in various prosodic issues in Czech and also in applications aiming at automated speech recognition.

REFERENCES

[1] Palková, Z. (1994), *Fonetika a fonologie češtiny,* Praha, Karolinum 1994.

[2] Janota, P. and Palková, Z. (1974), "The auditory evaluation of stress under the influence of context", *Phonetica Pragensia IV: Acta Universitas Carolinae,* Praha, pp. 29-59.

[3] Palková, Z. and Ptáček, M. (1994), "Ein Beitrag zur Intonation in der Diphonsynthese", *Phonetica Pragensia VIII: Acta Universitas Carolinae,* Praha, pp. 61-74.

[4] Palková, Z. and Ptáček, M. (1994), "Der Sprechtakt als eine rhythmische Einheit in der Diphonsynthese der tschechischen Sprache", *Speech Processing: 4th Czech-German Workshop,* Prague, p. 23.

[5] Palková, Z. (1987), "Intonatorische Merkmale in der Perzeption der Wortgrenzen im Satz", *Proceedings of the XIth International Congress of Phonetic Sciences,* vol. 1, Tallinn, pp. 296-299.