Elena KARNEVSKAYA
Mińsk

CURRENT PROBLEMS OF PROSODIC MODELLING FOR SPEECH SYNTHESIS

Results of speech synthesis quality evaluation tests, irrespective of the type of synthesis (formant, microwave, allophonic, etc.) elucidate a close association between the overall perceptual assessment of artificial speech and its prosodic characteristics. These results apparently testify to the essential role of prosody in speech communication and highlight a demand for further development of the prosodic component in the synthesis programmes. Among the properties of a synthetic utterance immediately dependent on the prosodic characteristics, one could list, in addition to and apart from those reflecting the commonly recognized linguistic functions of prosody, such as the intelligibility of synthetic speech, its stability in relation to noise, the degree of its naturalness and social-pragmatic acceptability.

The task being considered obviously involves a wide range of problems. These begin with an overall improvement of the linguistic base underlying the complex integrated mechanism of speech production in the "synthesis from text" programme (Lobanow 1987, Lobanov, Karnevskaya 1991, Lobanow 1991). The first requirement here is to determine the units of prosodic modelling as well as their internal organization and perceptual-acoustic discretization.

The prosodic model under discussion is part of a multilanguage synthesis programme (Karnevskaya 1987, 1996), and the solution of the aforementioned task is being sought in it with a view of the model's applicability to different languages. This means that the model aims at typological prosodic representation of the languages involved in the programme. The prosodic block of the latter is composed by three parametrical constituents – tonal, dynamic and temporal – each having an identical structure. They are actualized simultaneously, producing, as is the case in natural speech, an overall prosodic structure or contour. Its integrity is predetermined by the common segmental base for the parametrical constituents: they are all coextensive with an utterance or a part of it forming a separate "intonation-group", or "syntagm".

The contour is a multi-componential unit not only in the "vertical" aspect but "horizontally" too. The complexity of its linear organization derives from the possibility of breaking the syntagm down into elementary semantic blocks – accentual groups (units). As any hierarchically higher linguistic unit, the contour can be ultimately reduced to such a block that is an accentual group.

The unit of prosodic modelling having been defined, the next stage is to establish an inventory of contours that would correspond to the paradigm of prosodic distinctions in the given languages. Like all linguistic units, prosodic contours must be selected along functional criteria. Accordingly, the number of contours will be determined by the communicatively relevant prosodic contrasts discovered in the languages under investigation.

It must be noted, however, prior to the phonological task of establishing the contour types comes the fact of recognizing different prosodic patterns, or "figures", on the perceptual level as well as the possibility of analyzing their structural "content" into constituent features. A more accurate statement then would be that the principle of a contour identification is two-fold: phonetic and phonological.

As has been shown by experimental studies, languages display remarkable similarity in the repertoire and general shape of prosodic "forms". Phonetically similar patterns, furthermore, have similar semantic connotations.

This circumstance provides an objective basis for a typological approach to prosodic modelling in the present work. It leads, in effect, to the creation of a common inventory of contours for the languages being synthesized.

The structural description of a contour resembles in principle that of the "main" allophone of a phoneme: firstly, it is not restricted to the features relevant for paradigmatic differentiation but includes all the constitutive features of the unit; secondly, it represents the qualitative-quantitative characteristics of the contour in a "strong" position. The latter is identified here with the self-sufficiency of the contour's micro-context, leading to the relative independence of a contour from the macro-context. Regarded as meeting these requirements was a contour functioning in an expanded (having more than 2 words) syntagm (intonation-group) coextensive with a sentence either relatively isolated (a reply, an announcement, etc.) or initial in a text.

The accepted approach does not at all imply that the typological units coincide exactly in their acoustical-perceptual and semantic content. The programme takes account of the specific language features discovered not only on the phonetic level but in the semantic connotations, distribution and occurrence of the contours as well.

Needless to say, defining the ultimate prosodic parametrical values, or, in other words, building up the phonetic portrait of a contour, is the result of a complex, comprehensive analysis (perceptual and acoustic) of natural speech and a selection from the experimental data of such a realisation (or realisations) that can truly represent the given prosodic pattern.

The first criterion of the choice is the pattern's perceptual identification proved by a high degree of agreement among the listeners in the course of auditory testing. On the acoustic level, all (or nearly all) the functional segments of the contour realization viewed as a candidate for choice must reveal quantitative and/or qualitative differences with the other contours. Importantly, the comparison is two-sided: 1) with the members of the individual language paradigm and 2) between the typologically similar contours, occupying the same place in the interlanguage matrix in the programme.

Although some or even most of the specific traits thus revealed are not distinctive in the linguistic sense, they all contribute to the impression of prosodic "normativeness", which understandably is of special significance for speech synthesis. It prevents from ignoring subtle phonetic peculiarities and, as a result, brings synthetic speech nearer to the concrete natural language.

Improvement strategies for the prosodic model in the present programme concern all the components and constituents of the former. Further development particularly involves: 1) a critical overview of the inventory of contours both with regard to their total number, which must be limited yet sufficient for each selected language, and to their language-specific phonetic content and functional-distributional characteristics; 2) the adoption of a more flexible contour structure which would allow of elucidating various degrees of cohesion between syllable sequences within the contour; these reflect variation in the intra-clausal relationships, depending on the syntactical links of adjacent elements, the syllabic length of each accentual unit etc. 3) a modification of the linear structure with an aim of a more subtle representation of accentual groups, so as to bring into play, as much as possible, the contribution that each syllable within an accentual group makes in the realization of the overall pattern.

The second and the third points clearly concern the stage of the phonetic realization of a contour and are aimed at finding sort of a "compromise" between the continuity and discreteness of the latter. It should be noted hereby, that the feature "continuity" is inherent to the model due to the principle of hierarchy underlying the contour structure. According to this principle the phonetic characteristics of a microunit (accentual group, in this case) are determined by the type of the macro-unit, i.e. the contour incorporating it, on the one hand, and the position of the micro-unit in this larger unit, on the other. The discreteness of the contour, in its turn, being predetermined by its accentual-rhythmical segmentation, is obtained through a certain degree of prosodic completeness of the contour constituents, each of which presents, as a result, a relatively autonomous micro-structure (tonal, dynamic, rhythmical-temporal).

Since they are conditioned by the semantic relations within the intonation-group (syntagm), both the contour continuity and discreteness acquire a dynamic character in speech and their interaction leads to variation which manifests itself in the degree of the above-mentioned structural independence of the micro-units constituting the contour. This kind of variation is essential to the model because it adds the flexibility that the contour inevitably lacks if its identification is based on exclusively communicative-pragmatic factors. With a special mechanism of making the modifications of the above-said features feasible, the contour virtually becomes compatible with a multitude of utterances coinciding in the communicative-pragmatic but differing in the logical-semantic sense, i.e. in the syntactical relations and the distribution of informational weight between the elements of an utterance. The choice of an appropriate variant in a concrete situation is made possible by discriminating gradations of junctures – close, loose, neutral – between the accentual groups established in the course of the pre-acoustic text-processing procedure. In accordance with a complex accentual group functioning as a single prosodical-semantical block. When the juncture is loose, on the other hand, the deviation from the normal, i.e. neutral, type consists in that the contour becomes intrinsically split without losing, however, its semantic integrity. It means that despite some specific structural features enhancing the isolation of the accentual groups the contour is still perceived as one intonation-group. In other words, the juncture type assignment depends on the type of utterance stress (full, weak, partial, zero) which the word is entitled to in a concrete context, on the successive order of stress types and their general number in an intonation-group, as well as on some other factors including morphological and communicative-pragmatic. Alongside external or inter-accentual, junctures, the type of inter-accentual linking of the constituents is marked, too. An obvious distinction here is between a syllable and a word juncture.

The type of variation under discussion does not affect all the parametrical layers of a contour in the same way. The most significant changes are to be found in the timing (rhythmical) and tonal components. The particular modifications consist in smoothing or, vice versa, sharpening the duration contrasts between the marginal elements of the adjacent groups as well as in diminishing or increasing the degree of tonal completeness of the accentual groups. The criterion here is similarity between the ff. movement occurring within the given accentual unit and a pitch change identified on the level of perception with a “kinetic” tone. Another important feature is homogeneity/heterogeneity of the overall tonal pattern of the contour. In the second case the junctures between the adjacent semantic microunits within the intonation group are more likely to be perceived as loose.

In conclusion, it should be noted that the prosodic block improvement strategy presented in this paper implies a more sophisticated approach to the internal structure of a prosodic unit. The advantage of the suggested model is its flexibility: the model is oriented from the start to variation due to the instability of prosodic contour’s external (sound and syllabic) and internal (semantical-syntactical) form.

References

Karnevskaya E., 1987, *The Linguistic Aspect of Multi-Language Speech Synthesis.* Proceedings of the XI-th Congress of Phonetic Sciences. Tallin, Vol. 1. Pp. 98–102.

Karnevskaya E., 1996, *Prosodic modelling for speech synthesis: requirements and perspectives.* Описательная и сравнительная фонетика: теоретические и прикладные проблемы. – Мн: МГЛУ, с. 30–34.

Lobanov B., 1987,*The Phonemophon Text-to-Speech System.* Proc. of the XI International Congress of Phonetic Sciences, Tallin, Pp. 61–64.

Lobanov B., Karnevskaya E., 1991, *MW Speech Synthesis from Text.* Proc. of the XII International Congress of Phonetic Sciences, Aix-en-Provense, France, Pp. 406–409.

Lobanov B., 1991, *Microwave Speech Synthesis from Text.* Proc. of the 24 Fachkolloquim Informationstechnik, Dresden, Pp. 118–120.

AKTUALNE PROBLEMY MODELOWANIA PROZODYCZNEGO W BADANIACH NAD SYNTEZĄ MOWY

Streszczenie

W artykule przedstawiono uwagi dotyczące modelu prozodycznego, wypracowanego w ramach międzynarodowego programu badań nad syntezą mowy. Model ten ma mieć zastosowanie w różnych językach objętych tymże programem. Wyniki badań testów oceniających jakość syntezy mowy wskazują na bliski związek między ogólną oceną sztucznej mowy a jakością jej cech prozodycznych. Blok prozodyczny obejmuje trzy parametry: tonalny, dynamiczny i temporalny, z których każdy ma identyczną strukturę; aktualizowane jednocześnie, tworzą całościową strukturę prozodyczną, wielokonturową jednostkę, która może być rozkładana na grupy akcentowe. Na podstawie komunikatywnie relewantych kontrastów prozodycznych ustala się inwentarz konturów dla poszczególnych języków. Strukturalny opis konturu przypomina zasadniczo deskrypcję głównego allofonu fonemu. W programie uwzględniono nie tylko właściwe danemu językowi fonetyczne cechy konturów, ale również konotacje semantyczne, dystrybucję oraz frekwencję. Przyjęcie szerokiej koncepcji struktury konturu umożliwia przedstawienie różnych stopni spójności między sekwencjami sylab w obrębie konturu. Program umożliwia doskonalenie modelu prozodycznego i bardziej szczegółowe przedstawienie grupy akcentowej. Zaletą proponowanego modelu jest jego elastyczność, uwarunkowana zmiennym charakterem konturu prozodycznego, zarówno wewnętrznego, jak i zewnętrznego.