THE LINGUISTIC ASPECT OF MULTI-LANGUAGE SPEECH SYNTHESIS

YELENA KARNEVSKAYA

Minsk State Pedagogical Institute of Foreign Languages
Minsk, Byelorussia, USSR 220662

ABSTRACT

The present program adopts language-independent principles of modeling segmental and prosodic units of speech. Phonetic description of these units is based on experimental contrastive analyses of Russian, English, French and German phonetic characteristics and is carried out within a single normalized range of acoustic parameters.
The linguistic program includes rules for letter-to-phoneme conversion, phrasing and accent location rules, as well as algorithms for a prosodic contour choice and modification under varying contextual conditions.

INTRODUCTION

Rapid improvement of speech synthesis technology over the last two decades has resulted in the appearance of new programs permitting a wide range of user-specified applications. The general trend toward greater flexibility of synthesis systems is well seen through the growing interest in text-to-speech synthesis designs, and especially those handling a variety of languages /I/.
Multi-language systems apparently derive from programs suited to the needs of one particular language. Linguistically, this is justified by a universal, language-independent nature of phonetic categorization, which predetermines a largely universal character of speech synthesis as an analogue of natural spoken language. Thus any synthesis model will distinguish classes of phonemes and reflect coarticulation processes that sounds undergo in running speech; it is also bound to convey the polyparametric nature of speech prosody and take account of its linguistic uses relating in all languages to the communicative contents of an utterance.
Furthermore, articulatory and acoustic similarity of sound units belonging to identical classes in different languages - as a consequence of the phonological systems' typological similarities - suggests a possibility of applying a single descriptive apparatus for indentifying the phonetic features of the languages in creating the data bases for multi-language synthesis.
The above general prerequisites find ample explication in the current program which is based on a model, originally devised for the Russian speech synthesis. This model's applicability for multi-language purposes is due to the linguistically invariant principles underlying the design of its fundamental elements - portraits of phonemes and prosodemes /2/. Phonemes' acoustic portraits, in particular, incorporate a sufficient amount of parameters to convey exhaustive information about phonetically significant features of vowels and consonants of any language. The acoustic parameters, specifically, are presented as complexes of interacting targets and transitional functions whose values are determined by the unit's inherent properties, on the one hand, and the influence of its environment, on the other. Importantly, there are no constraints on a phoneme description either as regards parameter value modifications or the unit linear subsegmentation. It is clear that the portraits in question serve as a convenient tool for achieving allophonic output on the basis of phonemic input in conformity with the main principle of language units' actualization in speech.
In the same way, portraits of prosodemes are built in accordance with the assumed language-independent structure of an intonation-unit. They provide a sort of mould for embodying qualitative and quantitative properties of the selected patterns. Realization of the patterns is achieved by applying special rules for prosodic feature modifications depending on a number of previously defined variables.
However, multi-language orientation causes inevitable alterations of the original model, which essentially consist in adjusting the latter to the overall, broader program of which it becomes only a part. For the linguistic component this implies elaboration of a single classification matrix, which is optional in the sense that certain types or gradations may remain unoccupied in a concrete language, and at the level of acoustic analysis it suggests using absolute-relative characteristics, rather than purely relative, as preferable for the purposes being discussed. The main argument here is that an absolute-relative scale provides greater accuracy and precision in revealing inter-languages phonetic differences, especially in the case of typologically similar units.
Evidently, this kind of scale can be achieved only by normalizing data relating to different languages within a single range of parameter features: formant frequencies, amplitudes, pitch levels.
The idea of single acoustic space emerges in association with the processes of articulatory program shifting, commonly observed in the speech of bilinguals (multilinguals). It would seem that the conformity of the suggested approach to phenomena of natural speech gives ground for considering it valid.

I. Description of Phonetic Features

I.1. Material and Procedure

A normalized range of parameters has been obtained as a result of special contrastive studies in which Russian phonetic units were consistently compared to English, German and French ones.
Following the above assumptions we decided to have our test materials recorded by bilingual speakers in addition to having provided native speakers recordings.
The use of bilinguals, as an experimental method in a study of this kind, has got obvious advantages in that much of irrelevant acoustic variation is avoided, and search for interlanguage differences is thus facilitated. However, these strong points can hold only if the bilingual speaker's command of the second language pronunciation norms is really good, near to native, in fact, since phonetic interference otherwise pertinent may seriously invalidate the results.
With these requirements in mind test recordings were carried out by 3 bilingual speakers whose performance was assessed as normative (highly acceptable) by English, French and German native listeners, respectively.
The materials themselves were also constructed with a view of reducing, as much as possible, uncontrolled phonetic variability. The first set of utterances, for instance, were built in each pair of languages exclusively of the so-called interlanguage homophones, e.g. Klin(R.)- Clean(E.).
The words have been selected with a view of covering all permissible combinations of CV type in the languages being compared (as well as VC).
These words were grouped in three to produce nonsense utterances which were pronounced as declarative sentences of the type "John loved Mary".

I.2. Formant Characteristics

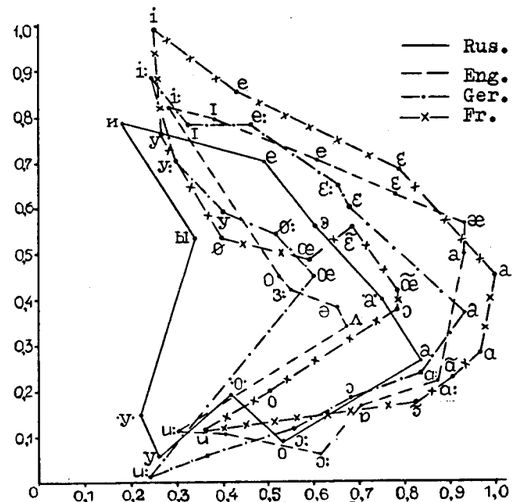

Fig. 1. Distribution of vowels on a normalized FI/ FII plane.

Table 1. CV coarticulation (F II).

V / C	u			a			i		
	Eng.	Ger.	Fr.	Eng.	Ger.	Fr.	Eng.	Ger.	Fr.
t	1,60	1,80	1,50	1,30	1,26	1,20	0,83	0,85	0,92
p	1,02	1,09	1,02	0,97	0,94	0,97	0,83	0,86	0,92
l	1,39	1,60	1,29	1,14	1,16	1,16	0,74	0,88	0,93

Through comparison of the distances between phonologically similar units in Russian and each of the other three languages frequency values of the first three formants were found for the vowels and sonorants, and the commonly recognized inter-language differences in the degree of CV, VC and CC coarticulation were evaluated. Some results of this study are shown in Fig. I and Table I above.

I.3. Prosodic Characteristics

The present model is based on the concept of prosodic contour as a major operational unit of non-segmental organization of speech. In view of the complex parametric structure of the contour a componental approach has been offered here, as in most current work, both to its analysis and synthesis. Obviously enough, the components into which the overall model is split are the fundamental frequency, intensity and duration contours. These can be taken as representing the pitch, accent and timing perceptible patterns of speech only if due notice is given to their close inter-

action in producing (and consequently,modeling) the intonational effects ultimately aimed at. Therefore,interaction of the components must become one of the underlying principles of the model.
Realization of this principle includes several aspects. For one,connection between the contours is ensured by their compatibility due to co-extensiveness with one and the same segmental base and identical internal structure: contour of any type is constituted by one or more (typically 2-3) accent-groups with the nuclear group as an obligatory element.
While the nuclear accent-group plays a special part in the characterization of all contours,its predominant role stands out most clearly in the F0 contour. The status of the latter in the overall prosodic model,in general,differs from that of the other two components and this functional inequality is another form of the contours' interaction.
Specifically,the total set of prosodic contours used in the program is determined by the number of tonal patterns that have been shown to be significantly contrasted. It needn't be argued that such a relationship is well in line with the widely accepted theories of utterance prosody. The implications involved here are that the role of intensity and duration modifications is confined to that of accompanying features contributing to the tonal pattern ample realization. However, these modifications are only partially controlled by F0 patterning. It is widely known that increases and decreases in intensity and duration of sounds are utilized by language in various other ways. Importantly,variations along these two parameters are rather more closely,by contrast with F0 changes,associated with the intrinsic properties of segmental units and such aspects of prosodic organization of speech as rhythm and stress,whose functions are distinctly different from those performed by F0 modulations.
The present program takes account of the observed peculiarities of prosodic parameters. It has been assumed that relevant information pertaining to the duration and intensity contours can be carried by segmental units,if their current temporal and dynamic characteristics are determined by special algorithms in which both segmental and prosodic variables are dealt with. The F0 model in its turn is not completely independent upon the characteristics of the segmental base upon which a contour is "superimposed",although intrinsic F0 influence is outside the scope of the reported work. On the other hand,vowel length type and,to some extent,consonant manner class have been considered as capable of altering the shape of F0 configurations,in English and German,in particular. Presence or absence of marginal syllables in an accent-group has also been taken into consideration. This limitation is but an exception from the general model which is independent of the given factor. Yet it cannot be ignored,e.g. in case of the Russian rising nuclear tone of the rising-falling configuration (/\ ): its falling element is accomplished on the post-nuclear syllables and in their absence the given configuration will take the shape of a steep wide rise.
The role of the above factors has been confirmed in a number of listening tests in which some synthetic realizations of F0 contours displayed a markedly lower percentage of correct identification both of the communicative meaning of the speech unit (in terms of such dichotomies as complete vs.incomplete,interrogative vs.declarative,neutral,calm vs.categoric,expressive,etc.) and the phonetic type of the tonal pattern (in terms of pitch-change directional types and pitch-level gradations). Perceptual "deficiency" of these contours clearly stemmed from insufficient duration of their segmental bases - an effect noted in numerous earlier writings. This difficulty is overcome by supplying multiple acoustic correlates to a single functional type of contour.
There is also positional and combinatory variation of tonal patterns. The former is achieved by assigning lower values to one or more F0 peaks of the contour,the pattern as such remaining unaltered due to a zonal nature of perceptible pitch categories.
The object of combinatory variation is to avoid monotony when two or more functionally identical contour types are demanded by the context. In this case the rules modulate the shape of the contour elements so that the resulting contour is slightly different both phonetically and semantically from the basic one and can be defined as its close synonym.
The above rules are preceded in the program by phonetic description of the basic patterns,established as a result of experimental investigation /3/. Some peculiarities of pitch contours in the languages under study are shown in Fig.2.
A separate algorithm for a contour choice is designed on the basis of distributional tendencies and semantic properties of the contours,such as,e.g. preferable use (in English) of a falling-rising pattern in an initial parenthetical phrase,or a tendency toward using directionally similar contours in adjacent non-final syntagms,etc.
The temporal and dynamic algorithms start with establishing inventories of durational and intensity allophones,respectively, in accordance with the adopted principles of their classification. More detailed account will be given in this paper of the duration rules.
In the suggested classification all allophones (of vowels and consonants,alike)are characterised by a single set of parameters,each having several discrete gradations. As a result,allophones differ in a combination of parameter features,the number of distinctions ranging from 1 to 6. The maximal figure corresponds to the number of factors regarded as potentially relevant: type of juncture on the syllable's left and right;immediate segmental environment of the given sound;the pitch pattern (tone) of the accent-unit;degree of prominence of the given syllable. Only some of the possible feature combinations are selected,the larger part having been excluded apriori as insignificant. Each allophone in each phoneme class is assigned a coefficient which is a ratio to the phoneme intrinsic duration. The latter was identified with the duration of a sound in an initial stressed syllable of a word.In determining phonemic duration it was important to bring out quantitative peculiarities of phonemes within a class, on the one hand,and to display interclass and interlanguage differences,on the other.
Analysis of the "minimal pairs" of allophones (those differing in one parametric feature) yielded quantitative evaluation of the effect produced by each of the factors considered in the study. One of the findings here was that the ratio values changed within a fairly wide range, e.g. from 15% to 35%,or from 50% to 75%, depending on the concomitant factors. Thus,the shortening effect of a voiceless stop upon a stressed vowel was twice as high in a phrase final position as compared to an initial accent-group; the lengthening effect of prepausal position upon a stressed vowel in a monosyllable is the greatest for a falling-rising nuclear tone,and so on. The conclusion to be made is that positing coefficients of increase of sound duration under the influence of separate factors is invalid unless all the co-occurring segmental and prosodic conditions are taken into account. The content of an allophone in the suggested classification is just such a complex of co-occurring conditions, thought to be relevant for determining segment duration. It would seem that this approach captures the non-independent character of the factors significant for modifying segment duration /4/.
Prior to the choice and realization of contours is determination of an intonation-group boundary location and placement of accents inside this unit.
An attempt has been made to express the syntactical-semantic segmentation markers of utterance (previously singled out and classified for each of the languages) in terms of morphological features of constituent words,their position,environment and potential semantic weight.

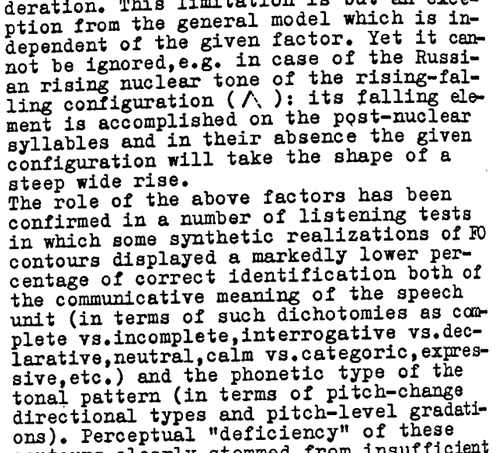

Fig.2. Phonetic realization of a declarative pitch contour.

CONCLUSIONS

The suggested language-independent classifications of parametric allophones and tonal contours as well as the absolute-relative methodology of phonetic description have proved applicable to multi-language synthesis. The present research has widely employed contrastive analysis of phonetic units as an indispensable stage towards speech synthesis. It must be stressed that analysis for synthesis is always analysis through synthesis as well, and this aspect is undoubtedly most interesting from the point of view of verifying the perceptual importance of the phonetic peculiarities revealed as a result of the analysis.
Further efforts are required to achieve greater formalization in the linguistic component of the program.

REFERENCES

/1/ Hertz S. "Multi-language Speech Synthesis - A Search for Synthesis Universals",J.Acoust.Soc.Am. 67, Suppl.1,1980, p.s. 39
/2/ Lobanov B. The Phonemophone Text-to-Speech System. In this volume.
/3/ Karnevskaya E.B., Lobanov B.M. Modeli sinteza melodicheskogo kontura russkih i angliyskih fraz. In: ARSO-82.- Kiev,1982.
Karnevskaya E.B. Tonalny kontur kak edinitsa prosodicheskoi organisatsii svyaznogo teksta. In: Sistemnye characteristiki ustnoi i pismennoi rechi. - Minsk, 1985
/4/ Klatt D. "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence", J.Acoust.Soc.Am. 59 (5), 1976, pp. 1208 - 1221.