THE PHONEMOPHONE TEXT-TO-SPEECH SYSTEM

Boris LOBANOV

Institute of Telecommunication
Leninsky pr., 113-20, MINSK, 220023, USSR

ABSTRACT

The paper report the results of the theoretical and experimental studies aimed at designing a universal text-to-speech synthesis model covering both the full range of intralanguage phenomena and applicable for multilanguage synthesis. The overall algorithm used in the text-to-speech synthesizer "Phonemophone" is described.

INTRODUCTION

The problem of converting a text into speech signal has been approached by several authors through the application of algorithmic synthesis or synthesis by rules [1,2] . A long list of rules and exceptions would be generally used, their number sometimes going up to thousand.
The main concern of the present study is to reduce the number of rules to a limited set of generalized categories capable of covering all the significant intralanguage phenomena and applicable at the same time to various languages. It is hoped that the lingual-acoustical model of the present work meets the above requirements.
The problem of speech synthesis can be split into two comparatively independent subproblems related to adequate synthesis of phonemic and prosodic structures of the text. At the acoustical level these subproblems correspond to the tasks of synthesizing the current formant parameters, on the one hand, and the current values of fundamental frequency, duration and intensity, on the other.
The present model of speech synthesis from text is based on two principles:
1. A limited set of acoustic invariant structures called portraits of phonemes and prosodems is used. They help to describe all the linguistically significant units of the phonemic and prosodic structures of the text.
2. A limited set of algorithmic rules is used to transform the portraits of phonemes and prosodemes into current acoustic features of running speech.

The exact number and the types of phoneme and prosodeme portraits are determined by available linguistic information about a given language. For instance, English phonemic units must be described by 20 vowel and 24 consonant portraits, in Russian 6 portraits of vowels and 36 portraits of consonants are required.
The exact number as well as the types of transformation rules for the portraits of phonemes and prosodemes rely on the up-to-date data in the fields of experimental phonetics, speech production and speech perception. Thus, for example,the number of transformation rules for the phoneme portraits must take account of the well-known effects of soundscoarticulation, reduction and assimilation.

1. FORMANT PORTRAITS OF PHONEMES

Phoneme and formant are fundamental notions of speech synthesis from text. A phoneme is an elementary and meaningful unit for any texts recording. The problem of transferring a written text into a phonemic one has already been algorithmically resolved for a number of languages and now doesn't present any difficulty. A formant, rather a formant parameter,is a universally unit for acoustic synthesis of any language sounds. A modern formant synthesizer can ensure the quality of sounds very near to natural.
Our model of speech synthesis employs the following set of operating formant parameters:
F1,F2,F3,Ff - frequencies of three voice formants and the generalized frequency of fricative formants (F-parameters); Av,An, Aa,Af - amplitudes of voice, nasal,aspirative and fricative formants (A-parameters). This set corresponds with the parallel-consequentive design of the formant synthesizer of speech signals.
The formant portrait of a phoneme is built on the 5 consequentive time segments: 0 - introductory, 1 - basic,2-3 - additional and 4 - the final segments In the phoneme portrait T0=T4=0, T1 always exceed zero, whereas T2, T3 can be equal or different from zero. Certain formant values of F- and A-parameters are given for each time segment. F-parameters are given for the 0-3 segments by three values of F, α, τ where F is inherent formant frequency, α - coarticulation coefficient, τ - formant transition duration. At 4th segment parameters are set at a value of τ^f only; A-parameters are set at segments 1-3 of values of A, τ^A and at segment 4 - by a value of τ^A only.
Thus, each phoneme portrait is described by 83 formant properties.
The portrait is graphically presented in Fig. 1.
The values of formant properties are established experimentally by analysing the behaviour of phonemes in natural speech. The minimal requirements to the experimental material are the following:
- each consonant is to precede and follow each vowel, as well as a pause;
- all the material is to be recorded by the same speaker.
Formant parameters and their properties are obtained by examining the sonograms of speech signal. The process of experimental analysis includes the phoneme fragments segmentation procedure, the normalization of the measured formant parameters, the determination of the inherent values of frequencies and coarticulation coefficients, the measurement of the formant transition duration.
As a result of the investigation phoneme portraits for Russian, Byelorussian, Ukrainian, Bulgarian, English, German and French were obtained and those later on proved valid in building the polylanguage system of speech synthesis.

2. TRANSFORMATION RULES FOR PHONEME PORTRAITS

The rules transforming the phoneme portraits into current values of formant frequencies are based on modelling allophonic variation in natural speech.
The major reasons of phonemes acoustic variation in connected speech are those of articulatory effects caused by coarticulation, reduction and assimilation.
Let's consider one of the most essential components of phonemes modification - that of coarticulation. [3] describes the model of coarticulation at the acoustic level. It has been shown in CV-syllable the formant frequency of the consonant F^{cv} can be expressed by inherent frequencies of the consonant F^c and of the vowel F^v by the equation:

$$F^{cv} = F^v + (1-\alpha^c)F^c \qquad (1)$$

where $0 \le \alpha^c \le 1$ is the consonant coarticulation coefficient. It is easy to show that the inherent consonant frequency and the coarticulation coefficient have the geometrical essense of consonant "focus" coordinates.

Fig. 2 illustrates the trajectories of formant frequency variation F2 (continuous lines) for the consonant /p/ and /t/ within syllables (PV, PA, PI).
The dotted lines indicate their continuation along the pause of the consonant with the point of intersection at the "focus" (point "a"). From the similarity of the triangles abc and cde it follows that

$$F^{cv} = \frac{\Delta 1}{\Delta 1+\Delta 2} F^v + \left(1 - \frac{\Delta 1}{\Delta 1+\Delta 2}\right)\varphi \qquad (2)$$

From equation (2) with (1) follows that the $\varphi = F^c$ and $\Delta 1/(\Delta 1+\Delta 2) = \alpha^c$.
Let's consider a more general case of a syllable containing more than one consonant, i.e. of the type C2 C1 V0, C3 C2 C1 V0 and the like. Spectrographic examination reveals in this case the dependence of the consonant formant frequency C2 not only on the vowel frequency V0 but on the consonant frequency C1. By analogy consonant formant frequency C3 is dependent on the frequencies C2, C1, V0 and so on.
To take this phenomenon into account the following recurrent equation was applied:

$$\begin{cases} F^{(1)cv} = \alpha^{(1)c} \cdot F^{(0)v} + (1-\alpha^{(1)c}) \cdot F^{(1)c} \\ F^{(2)cv} = \alpha^{(2)c} \cdot F^{(1)v} + (1-\alpha^{(2)c}) \cdot F^{(2)c} \\ \vdots \\ F^{(n)cv} = \alpha^{(n)c} \cdot F^{(n-1)cv} + (1-\alpha^{(n)c}) \cdot F^{(n)c} \end{cases} \qquad (3)$$

In the formula (3) the top indexes (n) denote the number of a consonant that comes in succession in a syllable beginning with a vowel marked (0).
Some coarticulation effects are also observed in vowel formant frequencies. For instance, in a particular environment F2 of vowels is considerably increased in the position before dental consonants and is reduced in bilabial environment. Modifications of vowel formant frequencies are calculated from the formula:

$$F^{vc} = \alpha^v(\gamma_1 F^{c1} + \gamma_2 F^{c2}) + (1-\alpha^v)F^v , \qquad (4)$$

where Fv, α^v are the inherent frequency and the coarticulation coefficient of a vowel phoneme; F^{c1}, F^{c2} - intrinsic frequencies of the adjacent consonants (both left and right); γ_1, γ_2 - weight factor. Algorithmic modification rules for the portraits of consonants and vowels affected by coarticulation are based on formulas (3), (4). The properties of Fc, α^c, Fv, α^v are taken from the tables of phoneme portraits. The phoneme portraits also carry the information required to simulate the effects of sound reduction and assimilation.

3. PROSODIC PORTRAITS

Speech prosodic features are intended as

a means of realising suprasegmental linguistic phenomena, those of stress and intonation in particular. In the proposed model speech prosodic units are hierarchically arranged in the following succession: syllable, word, accentual group, syntagm, phrase and, finally, phrasal unity (phonetic paragraph). Syntagm is the smallest independent prosodic unity. It is assumed that a limited set of syntagm portraits that can allow a relatively accurate description of speech prosodic features for reading any text can be established. The number of prosodic portraits is selected with a view to the need of realising such linguistic phenomena as communicative intention (statement or question, and the like), syntactical peculiarities (enumeration, contrast, etc), modality or emotion (command, surprise, etc). The number of prosodic portraits in a given language may go to several scores.

A syntagm prosodic portrait is established by a consequentive combination of prosodic portraits of the accentual groups, which fall into three types: initial, intermediate and final. These categories of accentual groups are unequal in their importance for the construction of the syntagm contour portrait. Major variety of contours is associated with the final accentual group carrying syntagmatic stress considerably smaller variety is observed within initial or intermediate accentual groups.

An accentual group corresponds to one or several words in a syntagm that are united into a single prosodic (pitch, rhythmic or dynamic) contour bound to one stress. When the number of accentual groups is the same as the number of words in a syntagm and an accentual group is identical to a phonetic word. But an accentual group may consist of more than one word, and in this case one of the words is made more prominent. The prominence of a word within the accentual group involves prominence of its word stress. Other words in an accentual group are less prominent and receive weak stress. It is more common that the main stress falls on the first word of the accentual group. This fact allows to define a very simple rule of determining the boundaries of accentual groups: the left boundary coincides with the beginning of the word having the main stress, the [illegible] comes in the beginning of the word of the next accentual group.

Pitch, rhythm and energy of the accentual group are established by the normalized values of fundamental frequency, duration and intensity over the three normalized time segments. They are nucleus, prenucleus and postnucleus. The nucleus of the accentual group is always the stressed syllable marked as the main stress. So the prenucleus and postnucleus are other phonemes of the accentual group that precede or follow the nucleus. The prosodic portraits of accentual groups are established with the help of tables of numbers having from two to four marked intervals over every time segment.

Fig. 3 is an illustration of the pitch component of the Russian prosodic portraits of the final accentual group for the six intonational types (statement, question, exclamation, enumeration, contrast, parenthesis). Pitch curves are outlined with the help of normalized coordinates "time – frequency". The normalized time interval (0–1/3) is a correlate of the prenucleus, (1/3–2/3) – of the nucleus, and the interval (2/3–1) is that of the postnucleus. The interval of the normalized fundamental frequency (0–1/3) indicates its low pitch level, (1/3–2/3) – the middle, and (2/3–1) – its high pitch level.

4. RULES OF PROSODIC PORTRAITS TRANSFORMATION

Pitch, rhythmic and dynamic features of accentual groups appear to be the minimal units that make up the intonation pattern of a syntagm, phrase and a text. Mention has already been made that the prosodic portrait of an accentual group is determined for the three normalized time segments – prenucleus, nucleus and postnucleus.

Within a natural text these segments may comprise a different number of phonemes, the latter in their turn may be of different duration in natural speech. Rules of prosodic portraits transformation are built with a view to these factors.

An example of pitch portrait transformation for the phrase 'Do it on Saturday' is presented in Fig. 4.

When an accentual group has no prenucleus or postnucleus part their function is attributed respectively to the left or right fragment of the nucleus vowel (1/3 of its duration). In case a syntagm includes more than 3 accentual groups the required number intermediate groups is added, and if the number of accentual groups is smaller than 3 the intermediate and secondly the initial accentual group are cut off. With regard to the position of a syntagm within a phrase or phrasal unity the portraits of accentual groups are subjected to further transformation that mainly consists in proportional modification of their scale.

5. TRANSFORMATION ALGORITHM TEXT-TO-SPEECH

The above described rules of establishing phonemic and prosodic portraits as well as the rules of transforming the current values of formant parameters from the ba-

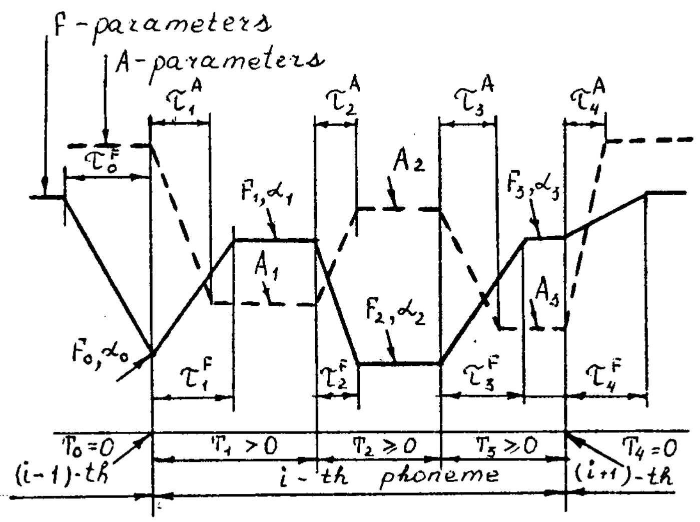

Fig. 1. Graphic presentation of a phoneme portrait

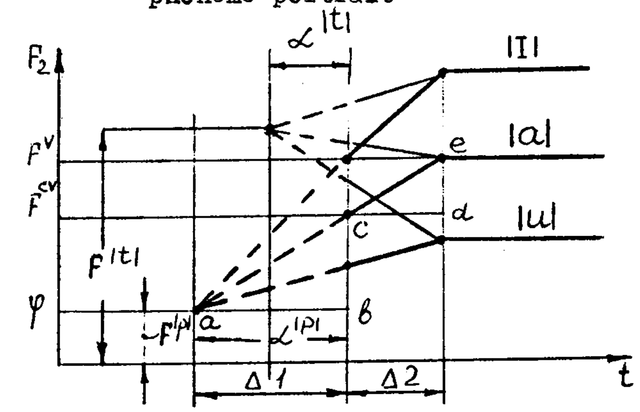

Fig. 2. Geometrical sence of F^c and α^c

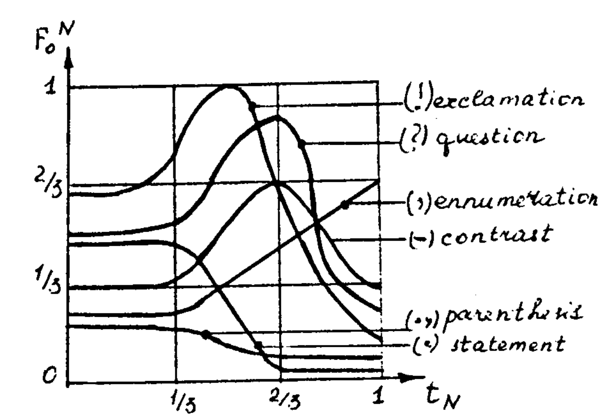

Fig. 3. Pitch component of the final accentual group prosodic portraits six intonational types (Russian)

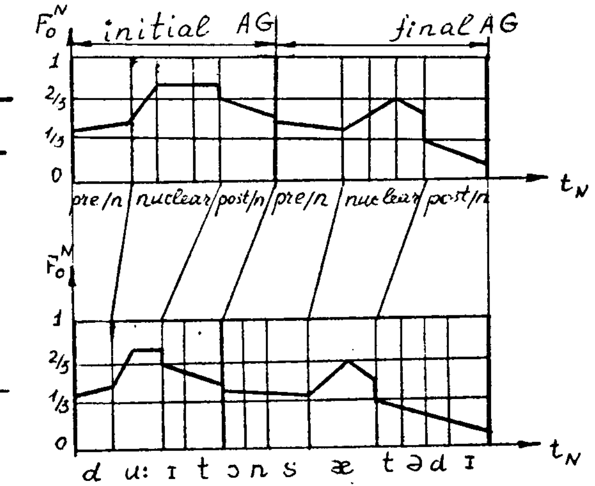

Fig. 4. Example of pitch portrait transformation

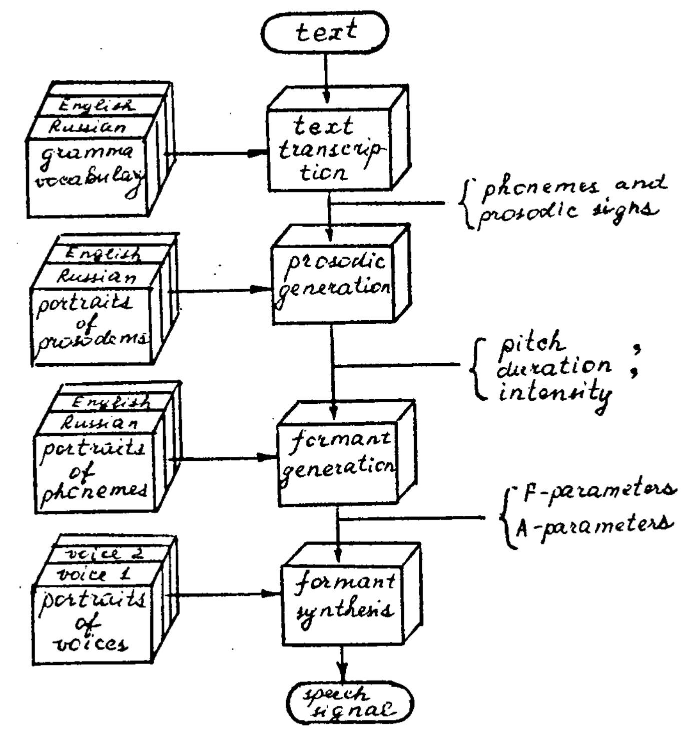

Fig. 5. Transformation text-to-speech algorhythm

sis of transformation algorythm text-to-speech in the Phonemophon system. The block-scheme of the algorythm is presented in Fig. 5.
At the first stage a written text is changed by certain rules, including a morphological vocabulary into a phonemic one which is provided with prosodic notation.
At the second stage prosodic parameters as based on the prosodic portraits and the rules of transforming frequency, duration and intensity are being generated. At the third stage formant parameters formed on the basis of phoneme portraits and the rules of transforming them into F-and A-parameters are being generated. From thus obtained sets of parameters many voices formant synthesis of speech signal is being performed.
The peculiarities of building phoneme and prosodeme portraits for multi-language speech synthesis and the rules of their transformation are described in [4] .

SUMMARY

The above presented strategy of speech synthesis from text formed a basis for compyling a series of Phonemophon devices. It has covered the distance from Phonemophon 1 to Phonemophon 5 since 1972 to 1987. On their basis since 1982 a mass production of speech synthesizers from text has been launched. The latest version of Phonemophon 5 is a single-card device built by digital microprocessors. It ensures a bylingual speech synthesis from text (Russian and English for instance) it is supplemented with controlled voice characteristics (3 male and 2 female) and with the controlled speech tempo. It is also well provided with the interface with a computer and telephone.

ACKNOWLEDGEMENTS

I gratefully record my thanks to Elena Karnevskaya for the beneficial cooperation in creating essential linguistic principles. I should also like to thank Michael Marchenkov and Irina Aksyutina for their participation in the constraction of the programmed model. I must thank Valeriya Afanasyeva and Yuri Zimitsky who have built the device itself. I am greatly obliged to my Institute's authorities and the colleagues of the laboratory who have readily supported this work.
And last not least I should like to thank Ludmila Leladze for the magnificent preparation of English version of this paper.

REFERENCES

1. D.H.Klat. The KLATTalk text-to-speech conversation system. In. Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. Paris, 1982, pp. 1589-1592.
2. Hertz S. "From Text to Speech with SRS". J. Acoust. Soc. Am. 72(4), 1982, pp. 1155-1170.
3. B.M.Lobanov. On the Acoustic Theory of Coarticulation and Reduction. IEEE Int. Conf. Acoust., Speech, Signal Processing. Paris, 1982, pp. 915-918.
4. E.B.Karnevskaya. The Linguistic aspect of multi-language speech synthesis. In this volume.