ABSTRACT
The paper reports 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 multilingual speech synthesis. The overall algorithm used in the text-to-speech synthesis "Phonemophone" is described.

INTRODUCTION

The problem of converting a text into speech 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 increasing up to thousands. The main concern of the present study is to reduce the number of rules to a 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 linguistic-acoustical coding of the present work meets the above requirements.

Speech synthesis can be split into two comparatively independent subproblems related to the acoustic synthesis of phonemes and prosodic structures of the text. At the acoustical level these subproblems correspond to the tasks of synthesizing the current formant frequency, duration and intensity, on the one hand, and the current values of fundamental frequencies of vowels, frequencies of fricative consonants, vowels, and fricatives (A-parameters). At the phonetic level the current formant frequency, duration and intensity, on the other hand, are obtained by examining the sonograms of speech. 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 the phoneme portraits for Russian, Byelorussian, Ukrainian, Bulgarian, English, German and French were obtained and those later approved valid in building the multilingual system of speech synthesis.

1. PHONEME 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 transforming a written text into a phoneme one has already been algorithmically solved for a number of languages and now doesn't present any difficulty. A formant is a universally used acoustic unit for various languages. A modern formant synthesizer can ensure the quality of vowels very near to natural.

The model of speech synthesis for the phoneme portraits must take account of the well-known effects of nonspeech articulation, reduction, and assimilation.

2. TRANSFORMATION RULES FOR PHONEME PORTRAITS

The rules transforming the phoneme portraits into current values of formant frequencies are based on modeling allophone variation in natural speech. 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 the phoneme portraits for Russian, Byelorussian, Ukrainian, Bulgarian, English, German and French were obtained and those later approved valid in building the multilingual system of speech synthesis.

3. PHONEMIC PORTRAITS

Speech prosodic features are intended as system of speech synthesis.
Fig. 1. Graphic presentation of a phoneme portrait

Fig. 2. Geometrical sense of \( p^0 \) and \( L^0 \)

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

Fig. 4. Example of pitch portrait transformation

Fig. 5. Transformation text-to-speech algorithm
sis of transformation algorithm text-to-speech in the Phonemophon system. The block-scheme of the algorithm 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 compiling 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 construction of the programmed model. I must thank Valeriya Afsanasyeva and Yuri Zinitsky who have built the device itself.

And last not least I should like to thank Ludmila Leladze for the magnificent preparation of English version of this paper.

REFERENCES