# MICROWAVE SPEECH SYNTHESIS FROM TEXT

B.LOBANOV AND E. KARNEVSKAYA

Institute of Engineering Cybernatics
Institute of Foreign Languages,Minsk,BSSR

ABSTRACT

The present work suggests a microwave method as a way to synthetic speech quality perfection. The paper deals with the basic principles of the adopted method and some problems arising in the course of its applicati on in a multilanguage program: choice of an invariant framework and specific types of microwaves ; the mechanism of microwave conca tenation and their modifi cations in accordance with linguistically significant prosodic changes.

## INTRODUCTION

Systems of speech synthesis from text today are general ly based on a formant signal method,as it permits a wide range of combinatory and positional modifications of the acoustic invariants representing the pho nemes of the language . It thus meets the requirements of the given type of speech synthesis , namely those of unrestricted vocabulary and sentence structure.Although modern formant synthesisers are capable of producing speech of a fairly high intelligibility and quality [1,2],much is left to be de sired.However,there is hard ly any possibility of a radical improvement at the present time . The reasons for it lie in the inherent deficiencies of the speech formation model being used and in particular, the latter´s inability to reflect voice individuality.This is largely because formant synthesisers neglect the interaction of the excitation source and the vocal tract (coupling effect) . Nor do they take account of the de pendence of the excitation pulse shape on the properties of the vocal tract modifications . As a result there still exist problems with the synthesis of a female voice as well as imitation of any definite voice.A way towards the solution of these difficulties, as it seems,is the use of speech segments as the basic elements of synthesis. The minimal units of synthesis in the present work are microwaves (henceforth, MW). They are elements of a natural speech signal coextensive with a F0 period.Ac tually,the use of microwaves for synthesis programmes was first proposed in [ 3]. In [4] this idea was successfully approbated in the system of diphone synthesis from text for male and female voices.Yet,there are quite a number of problems in MW synthesis that have not been solved so far (see Abstract). The present work being a multilanguage



programme lays special emphasis on finding language-invariant strategies and compiling language-specific MW sets.The number of micro waves in a set is ultimately determined by the phoneme inventory of the given language,phonetic distances between phonemes belonging to the same class and the difference in the degree of coarticulation between various types of sounds both within one language and across languages . The exact number and types of MW in each set , however,can only be defined experimentally.

## 2. GENERAL PRESENTATION

### 2.1.Microwave Phoneme Representation

Like in the formant synthesis,the basic principle of MW method is allophonic representation of the phone mes of the language , but unlike the former , there´s further disintegration of allophones into linear segments.Thus MW synthesis con sists essentially in obtaining adequate linear models of phoneme combinatory and positional realisations. Clearly there can be various degrees of discretisation both as regards the relevant list of allophones and their internal structure. The main argument for the validity of the MW sets selected for the present work is that they provide all significant variation of sounds in connected speech.For the Russian vowels, e.g., it is necessary first of all to distinguish between the soft and hard vowels:{A,O,E,U,I},on the one hand , and {´A,´O,´E,´I},on the other.It means that the target units are not phonemes in the strict sense of the word , but allophones viewed as sound types grouped on the grounds of non-functional,phonetic,identity . Each allophone of this kind,a higher-rank allophone,so to say,is represented horisontally by three successive segments : initial, mid and final.The segments, like boxes,are to be filled with appropiate microwaves, according to the modifications the given allophone (soundtype) undergoes under the influence of various ajacent sounds . In view of the accepted two-lewel allophonic representation the mid segment,i.e. the vowel stationary,was regarded in this paper as constant for all possible CV and VC combinations of a concrete higher-rank allophone . The choice of the MW type for the initial and final segments,as could be predicted from the results of formant analysis and synthethis , does not follow this principle:transitional microwaves vary in accordance with the ajacent consonant articulation place . The number of MW types then should cor respond to the number of consonant classes opposed by this feature.For Russian consonants, e,g.,we distinguish labial, dental,alveolar, velar and lateral places of articulation.
Allophonic variation of con conant phonemes in Russian and in English (as in other languages ) is caused both by the impact of the neighbouring consonants and vowels.The former may lead to noticeable qualitative chan ges ,e.g. the emergence of higher-rank allophones,such as the voiceless[r] in English. The latter is mainly confined to variations on the transitional segments. Thus,e.g.,we take three types of microwaves for initi al segments of hard conso-

nants: before [i,e], before [a] and before [o,u]. Clearly the three linear segments -are to be determined for each consonant allophone.

2.2. MW concatenation in the Speech Flow

MW concatenation at the stationary segment comes simply to their successive reading-outs. This procedure could be suitable for the transitional segments, too, if readings of several MWs for every type of transition had been preliminarily made. This can hardly be put into practice because of the amount of work needed for the preparation of the speech material and an excessive increase of the required memory volumes as well as the number of rules for the synthesis of transitional segments. There is an interesting possibility of avoiding these difficulties which is based on the use of the inertial properties of auditory perception.
Let us recall in this connection that the visual impression of a smooth replacement of slides can be achieved through a smooth decrease of the brightness ( down to zero) of one image and a simultaneous increase of the brightness(from zero up to the required degree ) of the other image , projected onto the same screen. Our research has shown that a similar effect of replacement is observed in sound perception. The auditory effect of smooth replacement is achieved by making an overlap interval between the contacting sounds during which a gradual amplitude decrease of sound 1 and a simultaneous amplitude increase of sound 2 takes place. The amplitude summing up in the field of the overlap leads to the appearance of a complex sound, perceived as a smooth transition from sound 1 to sound 2.

2.3.FO-Parameter Control.

The simplest method of controlling the fundamental frequency in the MW synthesis system is the following Let the initial MW have the duration TO which is chosen from the range of variations determined by prosodic rules:TOmin<TO<TOmax. As for a concrete TO value, it can be defined as a statistic mean value of the speaker's FO period used for the formation of the MW set. If the current TO = TO',the speech signal is formed by a simple repetition of the given microwave. When TO > TO',the MW repeated reading-out begins after the time interval TO-TO' and the interval itself is filled by zeroes. If TO < TO',the reading-out process stops at the moment = TO and a repeated MW read-out resumes. Experimental investigations of this control method have shown that it provides a sufficiently high quality of the synthesized sound, in particular, when TO>TO',with the interval TO - TO' not exceeding 30% max. of a period duration TO. When TO<TO there are no percertible distortions only if the end of the read-out falls at a MW value close to zero (10-20%% of MW amplitude). Otherwise, there's a clear sound distortion resembling nasalization. This unwanted effect can be removed by smoothing away the abrupt reading-out cessation process. It can be achieved by switching on an order 2 filter with the time constant equ. 0.25*TO at the moment the repeated read-out begins or before this moment (0.25TO) by multiplying the MW by a



smooth single function of the type y=exp(-t). The use of either of these methods for the case TO < TO'yields fairly good results . If TO > TO', the method of period zeroing can be applied provided TO~=0.7*TOmax.

3.IMPLEMENTATION

The algorithm implementing the above model consists of 10 blocks. The written text intended for synthesis is produced by the PC main program. This text, sentence by sentence, gets into block 1, in which sentences are segmented into intonation-groups and marked for stresses and melody. These procedures are performed in acccordance with definite rules varying from language to language. For every phoneme then in blocks 4,5 are calculated : the rhythmic ( sound duration), the dynamic (sound intensity) and the melodic (pitch) characteristics according to the rules specified for the languages . Futher on, in block 5, allophonic identification of the phonemes is carried out followed by the division of the allophones into linear segments. To each elementary segment corresponds a definite MW which is selected by block 6 out of the MW - set, determined for each language and type of voice. In block 7 modifications of the MW duration take place (i.e. of the FO period) in accordance with the information coming from block 4 , and in this way the tonal pattern of the synthetic speech is produced. Controlled by block 3, the duration of phoneme segments is defined in block 8 by means of a step-by-step reading of the required number of MWs. Finally in block 9 MW amplitude (intensity) is set out, while block 10 serves for smoothing the abrupt changes at the transitions from one MW type to another in the process of generating a continuous speech signal. Changing the voice type in MW synthesis is achieved by replacing or modifying the MW set . Passing over to another language inplies the replacement of the phonetic base rules and MW set

REFERENCES:

1.KLAT D.The Klattalk text-to-speech conversation system.Proc.IEEE ICASSP,Paris, 1982.
2.LOBANOV B. The Phonemaphone text-to-speech system. Proc. ICPhS,Tallinn,1987.
3.MOREL M. Synthese vokale par recordment de segment d'ooscillogrames Revue d'Acoustique,vol.14,no 56,1981.
4.HOMON C. Synthese par concatenation de formes d'ondes.
5.KARNEVSKAYA E. The linguistic principles of multi-language synthesis . Proc. ICPhS, Tallinn,1987.