



ON THE SPEAKING MODULE OF AN AUTOMATIC READING MACHINE

G. Olaszy (1), G. Gordos (2)

ABSTRACT

The speaking module -- called SCRIPTOVOX -- of the automatic Hungarian reading machine was developed in the years 1983--86 by a four-member research team (3) of electrical engineers.

The speaking module using the general purpose, programmable MEA 8000 type integrated circuit for speech generation, converts any Hungarian text -- given in ASCII codes -- into good quality speech.

(Only the automatic Hungarian text-to-speech (TTS) conversion is discussed below, the grapheme-to-ASCII character converter -- developed at the Institute for Computer Research, Budapest -- is not discussed here).

TEXT-TO-SPEECH CONVERSION

The primary requirement a text-to-speech converter system has to meet is that it should convert every character of a text in a given language (including not only letters but other characters as well) into control codes with the aid of which intelligible speech can be generated by a speech synthesizer. At the same time an important requirement is that it should recognize the different types of sentences (statements, questions, etc.). This recognition is the basis of the automatic generation of melody and rhythm. Last but not least, a fundamental requirement is the real time operation of conversion and speech generation .

THE SCRIPTOVOX SYSTEM

The conversion of ASCII characters of the text into synthesizer control codes is realised in the SCRIPTOVOX system

- (1) Institute of Linguistics, Academy of Sciences Budapest  
Pf.19. 1250 HUNGARY
- (2) University of Technology, Budapest XI.Stoczek u. 2.  
1111 HUNGARY
- (3) The team was headed by Dr. Géza Gordos (University of Technology, Budapest). The members of the team were Dr. Gábor Olaszy (Institute of Linguistics), György Podoletz (University of Technology) and György Takács (Research Institute of the Hungarian Post and Telecommunication).

in three steps.

1. Conversion of ASCII characters into "phoneme codes".
2. Conversion of phoneme codes into MEA control codes (speech frames) and their concatenation via a 255 frame data basis and a rule system.
3. Realisation of microintonation in some CV combination types as well as the melody patterns on sentence level and some other prosodic features.

#### Conversion of ASCII characters into phoneme codes

We use thirty-three phonemes (represented by numbers 1-33) for generating Hungarian speech. Only the short speech sounds are included among these thirty-three phonemes, the long ones are represented by doubling the phoneme code of the short counterpart. When processing the graphemes of the text -- stored in a 1 kbyte buffer ( $B_1$ ) -- into phoneme codes -- stored in a similar buffer ( $B_2$ ) -- three types of ASCII characters, are distinguished.

-- The first -- and simplest -- type comprises those characters with which a phoneme code can be associated directly in one step, e.g. A, A, O, V, F, H, etc.

-- The second type of characters cannot be converted directly into code numbers: their conversion requires an examination of the neighbouring characters. Examples of such characters are S, Z, C, T, etc. For instance, the letter S occurs in the combinations SZ, ZS, SSZ, ZZS, CS, CCS, denoting different sounds in each case.

-- The third group of ASCII characters includes numbers, abbreviations, and other symbols, like %, +, -, =, ", :, etc.

#### Phoneme code to MEA control code conversion

In the next step the program converts the contents of  $B_2$  into a series of speech frames which are stored in a 4 kbyte buffer ( $B_3$ ). For the conversion, a collection of speech frames (data base) and a  $33 \times 33 \times 6$  element concatenation matrix (rule system) is used. The data base consists of 225 different types of speech frames. The initial content of the 225 speech frames and the rules were defined in 1983 and have continuously been refined thereafter. The rule system includes rules for the concatenation of frames picked from the data base when converting the phoneme codes of  $B_2$  into a series of frames.

#### The rule system for the concatenation of speech frames

In order to get speech from the group of phoneme codes stored in  $B_1$  these codes must be converted into very many speech frames to be stored in buffer  $B_3$ . For the conversion 1200 rules -- incorporated in a  $33 \times 33 \times 6$  element matrix -- are used. The rule system works as follows:

- determines the necessary number and type of speech frames for every sound combination,
- performs certain sound replacements corresponding to assimilation phenomena,
- sets the proper duration of vowels depending on the surrounding sounds,
- imposes a falling intensity structure on the vowels,
- assimilates the steady state formant values of vowels to the surrounding sounds,
- realises the microintonation of certain CV combination
- builds in some prosodic features and variations in rhythm,
- decodes the punctuation marks.

After the conversion a monotonous version of the speech sequence that must be uttered for reading the text is present in B3 in the form of speech frames lacking data for melody generation.

#### Automatic generation of melody

To make speech more natural, melody patterns must be superimposed on the segmental realisation. In spite of the complicatedness of handling the pitch control in the speech frames of MEA 8000, a fully automatic melody generation was developed for the SCRIPTOVOX system. The melody is generated on a male voice timbre.

What are the elements of this melody generation?

1. Building microintonation into the appropriate speech frames.
2. Recognizing the articles and some conjunctions in the text and making them unstressed by reducing the pitch.
3. Recognizing comma(s) in the text and changing the intonation and rhythm before the comma(s).
4. Superimposing the intonation of declarative sentences characterized by a full stop at the end.
5. Superimposing the appropriate melody patterns on the various types of questions (question mark at the end). The types of questions distinguished for Hungarian are as follows.

Questions beginning with a Q-word (Mikor indulunk? 'When do we start?').

Questions in which the nucleus (the word questioned) has one syllable (e.g. Jó? 'OK?', A fagyj jő? 'Do you like the ice-cream?'). This type of question is called "one syllable question".

Questions in which the nucleus has two syllables (e.g. Hajó? 'A ship?' Ez egy hajó? 'Is this a ship?'). This type of questions is called "two syllable question".

Questions in which the nucleus has three or more syllables (e.g. Hajóval? 'By ship?', Elindultak már a gyerekek? 'Have the children started yet?'). This type of question is called "three or more syllable question".

## The perceptual examination of speech quality

The complete process of designing and constructing a TTS system has to end in a scientifically based perceptual examination of the speech quality. The acceptance of the system depends on the results of this examination.

A phonetically balanced speech material designed for the examination consisted of four groups of sound sequences: 30 syllables, 30 meaningless bisyllabic sequences, 30 mono- or polysyllabic words and 10 sentences. This material was recorded by a male announcer and by Scriptovox. Silent periods of 4--10s were left between the sound sequences. Thirty-six, 18 year old pupils and twenty adults took part in the test procedure. The natural speech material was given for two groups of 18 pupils in a classroom. One week later they listened to the synthesized material. The 20 adults listened to the material separately. They had to put down what they thought they heard. The results of the evaluation show that, from the word level upwards, the degree of the identification of the synthesized speech of the Scriptovox system is 98 % so it is quite close to that of natural speech. Further testing in everyday circumstances proved that the speech quality of Scriptovox is acceptable for use in industrial applications and in a reading machine as well.

## CONCLUSIONS

The SCRIPTOVOX text-to-speech system displays several differences when compared with conventional unlimited vocabulary speech synthesis systems. Its data base comprises only 225 four-byte data, each representing a speech frame with 8 to 64 ms duration. The rule system converting letters and other symbols into a concatenation of speech frames uses, at one point, a novel diad-like representation. Extensive experimentation was involved in formulating the rules of intonation for the various classes and subclasses of sentences. The memory requirement of the complete system does not exceed 12 Kbytes.

The SCRIPTOVOX system seems to accomplish a good compromise among low cost, high speech quality, fully automatic text-to-speech conversion (no need for diacritics or auxiliary symbols in the text), small memory requirement and a very low bitrate (approx. 100 byte/second).

## References

- GORDOS, G.-- TAKÁCS, Gy.: Digitális beszédfeldolgozás. Budapest, 1983.
- OLASZY, G.: A phonetically based data and rule system for the real time text to speech synthesis of Hungarian. Proceedings of the X<sup>th</sup> Int. Congr. Phon. Sci. UTTRECHT 1983, 225--230.