ON THE SPEAKING MODULE OF AN AUTOMATIC READING MACHINE

GABOR OLASZY

Institute of Linguistics
Hungarian Ac. of Sciences
Budapest 1250 Pf. 19
Hungary

GÉZA GORDOS

University of Technology
Budapest XI. Stoczek u.2
Hungary

ABSTRACT

The speaking module (Scriptovox) of the automatic Hungarian Reading Machine(RM) was developed in the years 1983--86 by a four-member research team of electrical engineers of the University of Technology, the Institute of Linguistics and the Research Institute of the Hungarian Post and Telecommunication. Scriptovox --using the MEA 8000 type integrated circuit for speech generation -- was developed for the fully automatic conversion of any Hungarian text into good quality speech in real time.

INTRODUCTION

The primary requirement a text to speech (TTS) 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 recognise 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 conversion of ASCII characters of the text into synthesizer control codes is realised in the Scriptovox system in 3 steps.

1. Conversion of ASCII characters into "phoneme codes".
2. Conversion of phoneme codes into MEA control codes (speech frames) and their concatenation.
3. Realisation of melody patterns by rewriting the pitch control bits of some speech frames of the group of frames concatenated in step 2.

LETTER TO PHONEME CODE CONVERSION

Thirty-three phonemes are used for generating Hungarian speech. Only the short versions of speech sounds are included among these thirty-three phonemes, the long versions are represented by doubling the phoneme code of the short counterpart. When processing the graphemes of the text into phoneme codes we distinguish three types of ASCII characters.

The first -- and simplest-- type comprises those characters with which a phoneme code can be associated directly in one step, e.g. 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 for 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.

This algorithm works as follows :

1. Setting of initial values .
2. Accepting ASCII codes into buffer-1 (B1) which is 1 kbyte large .
3. Identification of ASCII characters by scanning the text left to right.
4. If it is a special type of a letter then going to the rules for setting the appropriate phoneme code. Writing the phoneme code into buffer-2 (B2) which is 1 kbyte large.
5. If it is a number then going to the number routine where phoneme codes of the number will be inserted into B2.
6. If some other symbol then going to the lexicon and set the appropriate codes into B2.
7. Setting all other characters into B2.
8. Identifications of ASCII characters representing punctuation marks and setting the appropriate codes in B2.

Finally in B2 we find the original text in a form as if everything in it had been written using only letters. For example, the sentence MOST 12 ÓRA VAN. 'It's 12 o'clock now' takes the form MÖST TIZENKETTÖÖ OORA VAN.

PHONEME CODE - MEA CONTROL CODE CONVERSION

In the next step the program converts the content of B2 into a series of speech frames which are stored in a 4 kbyte buffer (B3). For the conversion, a collection of speech frames (225 different types) and a 33x33x6 element concatenation matrix (rule system) is used. The initial content of the speech frames of the data base 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 B2 into a series of frames. Choosing the appropriate elements of the 225-element data base every Hungarian sound and sound combination (VV,CV,VC,CC), as well as all the assimilations can be realised.

The rule system

In order to get speech from the group of phoneme codes stored in B1 these codes must be converted into very many speech frames to be stored in the buffer B3. This buffer size (4 kbyte) is enough for 40 s of speech in on conversion process. The rule system works as follows. In the rule matrix every row and column represents a phoneme code. The program -- before turning to the rule matrix-- makes a diad-like interpretation of the phoneme codes in B2 scanning it from left to right. So by the interpretation of consecutive phoneme codes a row and a column of the rule matrix are determined. This row and column pair points at a matrix entry, the contents of which are six bytes. These bytes represent the identifiers of speech frames that must be picked from the data base and placed into B3 one after the other. If the desired sound effect of a step during the conversion requires less then six speech frames ∅'s are inserted in the superfluous bytes. If during execution the program finds ∅'s it goes on to the next step. It should be noted that one complete sound combination is realised by the program totally after performing three steps: the step before the sound combination concerned in B2, the step of its own, and the next one. When the step by step conversion of phoneme codes is completed, B3 contains a series of speech frames of the text to be uttered. Sending these frames to the synthesizer a monotonous, robot-like speech will be produced. Thus the realisation of TTS conversion has been acomplished at the segmental level only.

Automatic generation of melody

To make speech more natural melody patterns must be superimpesed on the segmental realisation. In Scriptovox system a fully automatic melody generation of the mostly used types of patterns is working. What are the elements of this melody generation?

1. Building microintonation patterns into appropriate sound combinations.
2. Recognising the articles and some conjunctions in the text and making them unstressed.
3. Recognising comma(s) in the text and changing the intonation (and rhythm) before the comma(s).
4. Superimposing the intonation of declarative sentences characterised 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:

a) Questions beginning with Q-word.
b) Questions without Q-word -- further divided into three subcases(see below).

Microintonation

Quick variations in fundamental frequency independent of context and of the speakers' will are called microintonation. The variation ranges between 10--15 Hz inside a sound. It is built in the speech frames.

Articles and conjunctions

The system identifies the articles a,az 'the' and conjunctions és 'and', hogy'that'. To make these words less stressed the pitch is decreased by 8 Hz in them. In the article being at the very beginning of a sentence the pitch is decreased 16 Hz.

The interpretation of comma(s)

A comma in a written text corresponds to a change in the melody and rhythm of live speech. To implement these changes the proper place of comma(s) in B3 has to be marked. For this purpose a special frame with short duration and zero amplitude is inserted wherever there is a comma in the text. Then by scanning right to left the earliest 32ms long frame of the vowel immediately preceding the comma is searched for and its pitch is increased by 8 Hz. The search goes on to the left up to the next vowel and in its first 32ms long frame the pitch is increased by 4 Hz. In the frame following the comma the pitch is then restored to the former value.

The intonation of statements

The automatic generation of this type of pattern begins with the calculation of the length of the sentence. Three types of modifications are used in the algorithm, i.e. reducing the pich by 4 Hz, or by 8 Hz or reducing it in the last word of the sentence by an additional 4 Hz(see Table 1.) .

Table 1.

Sentence length categories	Minimum number of frames	Number of parts	Reduction in every part	Reduction in the last p add.	Reduction in the last word add.
very short	10	3	8Hz	-	-
short	20	4	4Hz	4Hz	-
normal	30	5	4Hz	4Hz	-
long	45	6	4Hz	4Hz	4Hz

The intonation of questions

The punctuation of questions is characterised by the use of question mark. In Hungarian several types of questions are used. To make an automatic realisation of questions, a multi-level algorithm has to be designed. Regularities lending themselves for algorithmic procedure, as well as unambiguous orthographic forms have been found for the following types of Hungarian questions.

Question beginning with Q word. The system recognises twenty-one Q words and implements the intonation pattern of Fig. 1a.

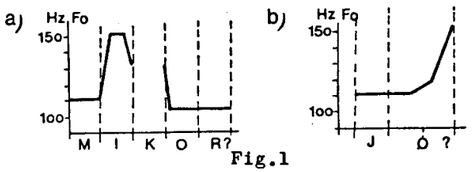

Fig.1

In the next step the algorithm of declarative sentences is applied in the rest of these questions.

Questions having no Q word are as follows: One syllable question, where the intonation peak must be at the very end of the vowel (Fig.1b).

Two syllable questions have the peak also in the last vowel but the pattern differs from the earlier one. The peak must be placed at the beginning of the vowel and the peak must be reduced towards the end of the same vowel. The duration of the peak must not exceed 30 ms. By this type of questions three types of word endings are distinguished (Fig.2). The intonation patterns differ radically in the way the pitch decreases at the end of the wovel.

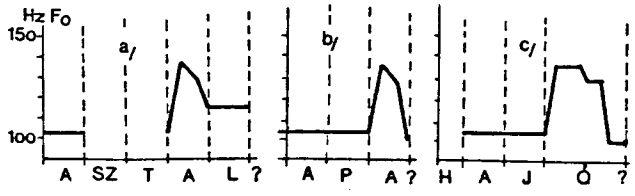

Fig. 2

Three or more syllable questions have the intonation peak in the last but one vowel followed by a decrease of pitch in the last one. The quality of the resulting pattern is further improved by distinguishing two subcases (Fig.3a,b). If the last but one vowel is long then the decrease of pitch has to commence already in the long vowel.

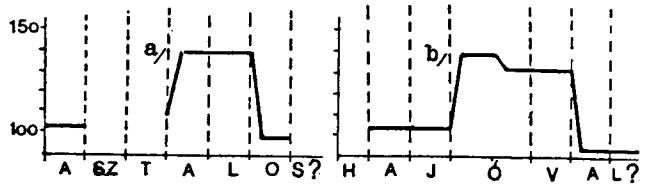

Fig. 3

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 phonetically balanced speech material for the test consisted of four groups of sound sequences: 30 syllables, 30 meaningless bisyllabic sequences, 30 one or polysyllabic words and 10 sentences. This material was recorded by a male announcer and by the Scriptovox. The natural speech material was given for two groups of 18 students (18 year old) and one week later they listened to the synthesized material. They had to put down what they thought they heard. The score for words was 84% for both the natural and synthesized items, as to the sentences, 99% for the natural and 98% for the synthesized was obtained.

CONCLUSIONS

The Scriptovox TTS system displays several differences when compared with conventional unlimited vocabulary speech synthesis systems. Its data base comprises only 225 speech frames (1 kbyte). The rule system converting letters and other symbols of the text into a concatenation of speech frames uses, at one point, a novel diad-like representation. Extensive experimentation was involved in formulating the fully automatic generation of rules of intonation for the various classes and subclasses of sentences. The Scriptovox system seems to accomplish a good compromise among low cost, high speech quality, fully automatic TTS (no need for diacritics or auxiliary symbols in the text), small memory requirement (12 kbyte) and very low bitrate (approx. 100 bytes/second).

REFERENCES

[1] MEA 8000 voice synthesizer. Philips Technical Publ. 101. 1983, Netherland

[2] Gordos G.--Takács Gy.:Digitális beszédfeldolgozás. Müszaki K. 1983, Budapest

[3] Olaszy G.:A magyar beszéd leggyakoribb hangsorépitő elemeinek szerkezete és szintézise. NyÉrt. 121. 1985, Budapest

[4] Olaszy G.:A phonetically based data and rule system for the real time text to speech synthesis of Hungarian. X-th Int.Cong.of Phon.Sciences, Utrecht 1983, Abstracts 398.