A Phonetically Based Data and Rule System for the Real-Time Text to Speech Synthesis of Hungarian

G. Olaszy
Budapest, Hungary

1. Introduction

Synthetic speech becomes more and more the focus of scientific, industrial and other applications. Speech synthesis by rule is a language-oriented task that means that the acoustical structure and the rules for building speech sounds, sound combinations and longer building elements have to be researched for every specific language.

The research of formant synthesis of Hungarian by rule has been done at the Institute of Linguistics in Hungary since 1979. During this research work we used a self developed analysis by synthesis method (Fig. 1.) to establish the data of the frequency, intensity and time structure of Hungarian speech sounds, sound combinations and longer sequences in such a form that the results could be used for the automatic building of words, and sentences by a computer and could be converted into speech by a formant synthesizer. For the analysis we used a Sound Spectrograph 700 and an Intensity Meter IM360, for the synthesis an OVE III Speech Synthesizer controlled by a PDP 11/34 computer.

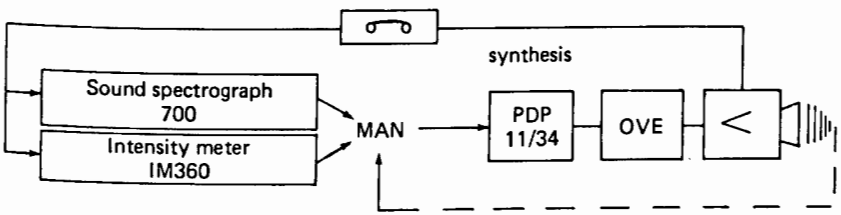

Fig. 1. Schematic diagram of the analysis-by-synthesis method used.

The software of the above mentioned synthesizing system was developed and made by Gabor Kiss (Kiss and Olaszy 1982).

2. Results

Our latest result (April 1982) is a *Hungarian speaking real-time text to speech synthesizing system* (named UNIVOICE), by which we can convert any kind of text into speech in real time. The system works without any vocabulary. It

can be used from a typewriter keyboard. The typed text will be spoken by the synthesizer immediately after giving the full stop (or !, ?) at the end of the sound sequence. Keyboard operations can be replaced by an ASCII code stream of the letters of the required text. Using this form of operation the system 'utters' the previously coded text automatically. As far as we know this is the first real-time text to speech synthesizing system that accepts Hungarian spelling rules and converts written text immediately into speech.

In order to determine the intelligibility of the system's output, perception tests were carried out by the author (see Gósy and Olaszy 1983). On the basis of test material collected by the linguist-phonetician M. Gósy, 70 subjects were asked to test syllable, word and sentence size units generated by means of UNIVOICE system. According to the test results the synthetic speech proved to be well understandable.

In this paper I give a short description of the data base of the UNIVOICE system and some rules used to build up the speech from the elements of the data base.

Synthesis by rule demands a data base that contains the necessary building elements of a language to be synthesized, and a computer program that can handle the data base according to the rules given.

The aim of our research work was to create a real-time text to speech system for Hungarian. Practically we had to place the data base and the program as well in the central memory of the computer. It means that we had only limited place for both of them (our PDP has 28 K word memory). Thus attempts had to be made to find the minimal number of building elements of the Hungarian language and the optimal number of speech sound parts. We can do this minimalisation and optimalisation only when we know the exact acoustical structure of the language examined and the technical operation of the synthesizer we use. This implies that - on the one hand - research had to be done to discover the frequency, intensity and time structure of the speech sounds, inherent sound parts, sound combinations, transition phases etc. in Hungarian. On the other hand specific technical and physical knowledge was necessary for finding how the acoustical data of the human speech can be used by an electrical system of a limited scope. Man produces speech with a biological system, we have to do it with a technical one.

3. Discussion

For Hungarian we found that the minimalised and optimalised data base of the language for OVE III contains 370 speech sound elements. This data base was developed in 1980-81 (Olaszy 1981, 1982a). The 370 elements are not speech sounds or sound combinations but they are speech sound parts.

Minimalisation of this data base implies that one sound part (element) can have the function of representing not only one speech sound but provides information at all places in sound sequences where the acoustical structure of this sound part meets the phonetically necessary requirements. This results in

sound parts that can be used at several places of the synthesized sound sequence and of course, there are those as well that can be used in the building process at only one position. Doing this process several times the number of building elements in the data base can be still further decreased. Finally we get the minimal number of building elements (in our case 370) that are necessary to build up any kind of words, sentences etc.

Optimalisation means that one has to find the optimal number of sound parts in the synthesis of a speech sound or sound combination. The more sound parts are used the better the sound quality is, and consequently the greater the memory demand. For example we can build a [b] sound from two, three, four or more sound elements. If we do it from two, the quality of the sound is bad, if we do it from three is becomes better and so on.

In the UNIVOICE system one speech sound and its transient phase consists of 3 or 4 sound parts. These sound parts are enough to realise the frequency, intensity and inherent time structure of the Hungarian speech sounds and sound combinations. One sound part is built from 1-5 microelements having the duration 4-50 ms. In one microelement the frequency and intensity data are constant. We can realise the formant and the intensity movements by making changes in the frequency and intensity data step by step in the microelements.

Using the UNIVOICE synthesizing system one can generate Hungarian speech sounds, syllables, words, words having no sense (for example for medical purposes), sentences and longer sequences as well. Non-Hungarian speech can be generated as well (English, German, Dutch, Finnish etc.) if we write the text phonetically using Hungarian letters. Of course the sound of any non-Hungarian language will be a little Hungarian-like because the UNIVOICE uses Hungarian phonemes only.

Hence by the synthesis process nearly every sound element of the data base can be linked to any other depending on the written text. This kind of operation demands that the 370 speech sound elements had to be planned in the way that if any of them comes into contact with an other – by the building of speech – the acoustical connection of them would be smooth without any transients, formant frequency shifts.

This data base for the real-time text to speech synthesis of Hungarian was developed for the OVE III speech synthesizer but it can be adapted for other formant synthesizers (for example MEA 8000) as well.

The program of the UNIVOICE system was developed and written by Gabor Kiss. How does it work? If we type a text on the keyboard of the terminal the program converts the letters to phonemes and phoneme combinations, after that it finds out which sound elements – from the 370 – are necessary to build the text, picks these elements from the data base and puts them into the appropriate order, determines the necessary melody pattern according to the punctuation marks typed at the end of the text and finally sends this data group to the input of the synthesizer.

The first demonstration of the UNIVOICE occurred at the 8th Collo-

quium on Acoustics, Budapest on the 6th of May 1982, where a 3 minute synthesized Hungarian speech was played to the audience (Olaszy 1982b)

A later version of UNIVOICE made it possible to make changes in the time structure of the typed text, melody patterns can be added at will and the intensity structure can be varied as well. For these changes the user only has to give some commands containing the data of the required time, melody or intensity structure.

4. Conclusion

Summarizing we can state that apart from a theoretical data base and rule system, a practical working model has been developed for synthesizing Hungarian of a good quality. The elasticity and speed of the system makes it useful for various industrial and other purposes.

References

Flanagan, J.L. and Rabiner, L.R. (1973). *Speech Synthesis*. Strassburg.

Gósy, M. and Olaszy, G. (1983). *The Perception of Machine Voice*. (Examination of the UNIVOICE, Hungarian speaking real-time text to speech synthesizing system). Nyelvtudományi kozlemények.

Kiss, G. and Olaszy, G. (1982). An Interactive Speech Synthesizing System with a Computer Controlled OVE III. *Hungarian Papers in Phonetics*, **10**, 21-45.

Olaszy, G. (1981). Preparation of Computer Formant Synthesis of Sound Sequences. *Hungarian Papers in Phonetics*, **8**, 147-59.

Olaszy, G. (1982*a*). The Analysis and Synthesis of the Hungarian Consonants and the Consonant-Vowel Combination Types. *Hungarian Papers in Phonetics*, **10**, 46-82.

Olaszy, G. (1982*b*). Some Rules for the Formant Synthesis of Hungarian. *8th Colloquium on Acoustics*, Budapest, Lectures, 204-10.