A HARDWARE-SOFTWARE SYSTEM FOR DESIGNING HIGH-QUALITY SPEECH COMPILING SYNTHESIZERS

ALEXANDRE GORODNIKOV

All-Union Research Institute
for TV and Radio Broadcasting
Moscow, USSR 123298

MEELIS MIHKLA

TOOMAS TAGO

Computer Design Office, Institute of
Cybernetics, Tallinn Estonia, USSR 200104

ABSTRACT

Nowadays, the highest possible phonetic quality of synthetic speech can be provided by compiling speech synthesizers. Proposed is an appropriate hardware-software system for their computer simulation and design. As an example of practical implementation, basic parameters of a high-quality speech synthesizer of the "speaking clock" type to be used in radio broadcasting are presented.

INTRODUCTION

Of different types of synthesizers available, it is the so-called compiling synthesizer which guarantees the highest possible quality of synthetic speech, and, consequently, boasts the greatest versatility. The synthesizer is based on a solid-state memory containing speech signals in a digital form. The set of signals consists of specially selected speech elements like phrases, words, syllables, or coarticulation units which, being read out from the memory in a preset order, permit to synthesize a certain number of utterances.

Designing a compiling synthesizer, the key problem is how to compromise among different and even somewhat antagonistic technical requirements, such as the quality of synthetic speech, the volume of the vocabulary, the complexity of the hardware part, dimensions, weight and cost. To provide an effective solution to the above problem, we have developed a hardware-software system that serves well for both research purposes and practical applications in creating compiling synthesizers. The system's hardware also includes a compiling synthesizer of the "speaking clock" type for high-quality speech synthesis.

THE HARDWARE-SOFTWARE SYSTEM

The system is based on a minicomputer EC-1010, operating together with 12-bit A/D and D/A converters, a bank of filters, a tape recorder, and other peripheral equipment. The system's features include: digital input of a speech signal, extraction of the synthesizer vocabulary units from a speech signal, computer simulation of the synthesizer operation algorithms, comparison of different methods of speech signal coding and redundancy reduction, objective analysis and comparison of prosodic characteristics and coarticulation joints of synthesized phrases. It is also possible to prepare and store in the solid-state memory bulks of labelled digital data and to check by listening the acoustic quality of synthetic speech. Sampling frequency of the speech signal input can be - depending on the application - 10,16 or 20 kHz. A segmentation program makes it possible to extract the wanted sentence, word, or syllable from a continuous speech signal. Thus derived speech elements are stored in a database on disks. The next step consists in the analysis and optimization of the vocabulary by means of synthesis. The prosody of a synthesized sentence and the intensity of the speech signal are compared to the corresponding parameters of an originally spoken sentence. According to the context of the sentence, the database is searched for speech elements whose main pitch contour and intensity most closely resemble those of the original sentence.

WORD SELECTION FOR THE SYNTHESIZER VOCABULARY

The highest possible quality of synthetic speech can be achieved in case the vocabulary consists of words and phrases. However, this requires a large-capacity solid-state memory, otherwise the synthesizer shall have a rather limited vocabulary. As an example, we may consider the vocabulary of a high-quality speech compiling synthesizer to be used for time announcement in radio broadcasting (the so-called "speaking clock"). The general structure of the Russian time announcement is as follows: "Moscow time is ... 10 x (hours), 1 x(hours)... 10 x (minutes), 1 x (minutes)" or "It is noon/ midnight in Moscow". Thus, in order to announce time with a minute's precision round the clock one would need 1440 announcements, each structured according to the above pattern and being 4-6 words long. The entire file of speech units used for time announcements would comprise 8592 words with a total duration of 183 min. Obviously, the vocabulary of a "speaking clock" should be considerably smaller in order to provide both a tolerable degree of complexity and a reasonable cost of the synthesizer.

A prosodic analysis of the original time announcemets carried out by means of our hardware-software system showed that abrupt changes in the pitch contour are observed mainly in the middle of the sentence in the words "time" and "(10 x) hours, (1 x)hours", where the pitch rises at the end of the word, and also in the sentence-final position in the words "(10 x)minutes, (1 x)minutes" where the pitch falls. The words carrying quantitative temporal information (i.e. numerals) can be divided into stressed and unstressed ones. In long words, however, changes in the pitch and signal intensity are relatively small, therefore the stressed vs. unstressed dichotomy is not worthwhile in this case. The above findings, alongside with the fact that most of the words display a high repetition rate across different announcements enabled us to considerably reduce and optimize the synthesizer vocabulary. The resulting vocabulary for round-the-clock time announcement service in Russian comprises 43 words with a total duration of 293 sec.

THE COMPILING SYNTHESIZER OF THE "SPEAKING CLOCK" TYPE

Prior to its practical implementation, the "speaking clock" design was simulated and optimized by means of our hardware-software system. The high acoustic quality of the announcements was achieved by 12-bit digital speech conversion with the 16 kHz sampling frequency. In order to economize the solid-state memory storage capacity, speech signals were DPCM-coded. The data-transmission rate was therewith 128 kbit/sec and speech signals were digitally encoded in the format of 8 bits per sample.

Figure 1 represents the block diagram of the compiling synthesizer. There are four main units: an electronic clock and a keyboard controller based on a one-chip microcomputer, a control and display panel, a CPU, and a solid-state memory. The overall dimensions are 475 x 280 x x 440 mm, the power consumption is 60 W.

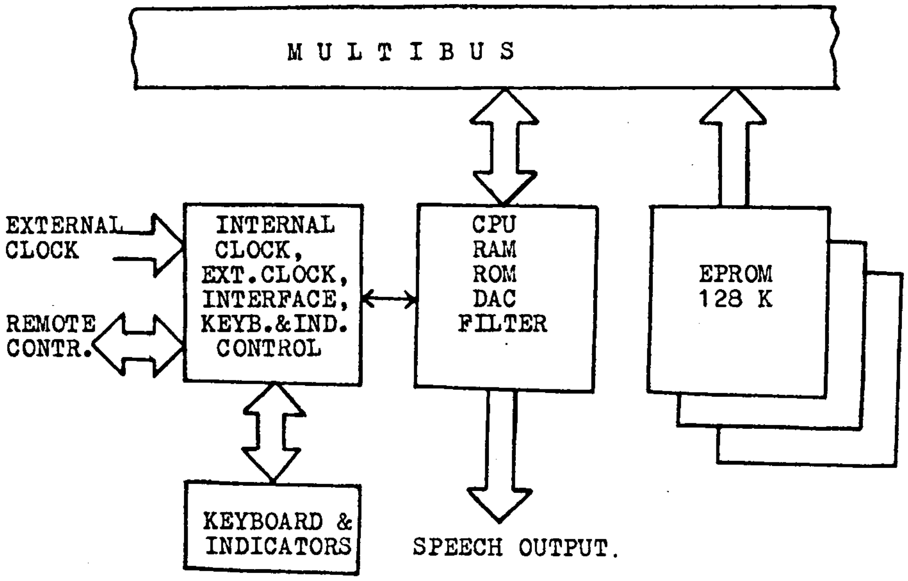

Fig. 1. The structured scheme of the compiling synthesizer.