5e CONGRÈS INTERNATIONAL D'ACOUSTIQUE
LIÈGE 7-14 SEPTEMBRE 1965

ON LINE COMPUTING IN SPEECH RESEARCH

P. B. Denes

Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey, USA

Over the last 5 to 10 years the digital computer has found many useful applications in speech research. In normal computer operations, however, the experimenter is somewhat remote from the process he is simulating because he must hand his program to an operator and some time elapses before he receives the results, in the form of a printout or a tape recording of the computer processed speech. More recently, "on-line" use of computers has become possible. In on-line work the experimenter remains in closer contact with the speech process he is investigating because he operates the computer himself, he can adjust selected program parameters during the processing, and observes the results in real-time. The purpose of this paper is to discuss why the closer man-computer coupling achieved in on-line work is of value in speech research, what its basic requirements are, and to describe one way of implementing these requirements.

Until recently it was thought that unique relationships existed between the phonemes and words of the language and specific acoustic features of the speech wave. Much time and effort was spent, therefore, on improving our knowledge of the acoustics of speech. As a result, we now have sophisticated ways of measuring the acoustic features of speech, and also synthesizers, which, when controlled in sufficient detail, will produce artificial speech that is very similar to the genuine, humanly-produced article. However, such synthesizers will generate good speech only if they are specially adjusted for the particular sentence to be produced. If the same acoustic features are used for the same phonemes in other sentences the speech may become unintelligible. Such and similar evidence made us realize that the acoustic features associated with any one phoneme are highly variable. They vary as a function of a great number of articulatory, linquistic and semantic factors, all of which I shall collectively refer to as contextual factors, and many of these are themselves interdependent. Useful as our extensive knowledge of speech acoustics is, we now know that the key problem in speech research is the understanding of how the acoustic features of any one phoneme are influenced by context.

The building and testing of models of the human speech mechanism has long been used as a powerful method for speech research. Electronic models using analog circuitry were relatively easy to build as long as they represented only the acoustic process of speech generation or analysis. Analog models which incorporate even the simpler contextual factors are much more difficult to realize and were not really practical until the digital computer became available for speech research. For example, in speech synthesis, it was not until Kelly and Gerstman[1] programmed an IBM 7090 to say "To be or not to be..." and sing "Daisy, Daisy,..." that the so-called synthesis-by-rule methods[2], which represented much of what we knew of contextual influences at that time, could be put to a practical test.

[1] J. L. Kelly, Jr. and L. J. Gerstman, "An artificial talker driven from a phonetic input", J. Acoust. Soc. Am. 33, 835 (A), 1961.

[2] F. Ingermann, "Speech synthesis by rule", J. Acoust. Soc. Am. 29, 1255 (A), (1957).

There is no doubt then, that computer methods are of great help in speech research. Yet, they have certain disadvantages compared with the earlier analog methods. In analog systems, the experimenter can readily change one parameter or another and immediately observe the corresponding change. In an analog synthesizer for example, he can control the frequency of a formant and hear the change in quality of the resulting sound as he turns the control knob. This close coupling with his model gives the experimenter a feel for the properties of his model, a greater insight into its operation than is possible by conventional computer methods, where several hours may elapse between adjusting the parameters of a model and being able to observe the results of the change. The advantages of close man-model coupling of analog systems can be combined with the great flexibility and logical power of computer methods by on-line computer simulation of speech processes. The rest of this paper discusses the basic requirements for on-line operation, and describes details of one such system and the work being done with it.

On-line processing requires (1) that the program run in real-time, (2) that the experimenter has means of continuously controlling selected parameters of the simulated model, and (3) that effective displays of the computed output of the model, in the form of an acoustic output or cathode ray tube displays of speech wave shape or spectrum available.

A DDP 24 digital computer with a 12,000 word, 24-bit memory, a 5 microsecond memory reference cycle, 3 index registers and very flexible input output facilities capable of transmitting up to about 350,000 12-bit samples per second was selected for our on-line operation. Let us see how such a computer fulfills the three basic requirements for successful on-line operation that were outlined in the previous paragraph.

The first requirement was real-time computation. The DDP 24 is a fast machine capable of executing about 100,000 instructions per second. Three index registers instead of just one, as is more usual in smaller machines, also help in the rapid execution of programs. However, computer simulation of the initial spectral analysis and of the final synthesis needed in most speech work is very time consuming. Even the relatively fast DDP 24 would not be able to perform them in real-time in addition to the computation of contextual influences on acoustic features. It was arranged therefore not to compute the final synthesis of the speech waves but instead make the computer only control the parameters of a conventional analog synthesizer attached to it. Similarly, for speech analysis, the computer will accept as input not only the original speech wave itself but also the multiplexed output of a bank of 60 filters to which the speech wave input has first been applied. The general lay-out is shown in Fig. 1. This mixed analog - digital processing arrangement in no way conflicts with the advantages expected from digital operation. As stated earlier, the flexibility of computer processing was required mainly for the logical manipulation of formants and other acoustic features as a function of a variety of contextual factors: efficient analog methods already exist for spectral analysis and for synthesis, and they may as well be used to speed up the overall operation of computer simulation.

The second requirement for effective on-line operation was that the experimenter should have immediate control of selected parameters of the computation. This feature is provided in our system by one or more control knobs, shown in Fig. 2. These knobs simply vary a DC voltage proportionally to their angular position. This DC voltage is sampled at frequent intervals, say every 10 msecs. Its value is converted into digital form, and then read into the computer where the program can use it for adjusting its operation. Our present installation has two independent knobs although there would be no difficulty in having a larger number.

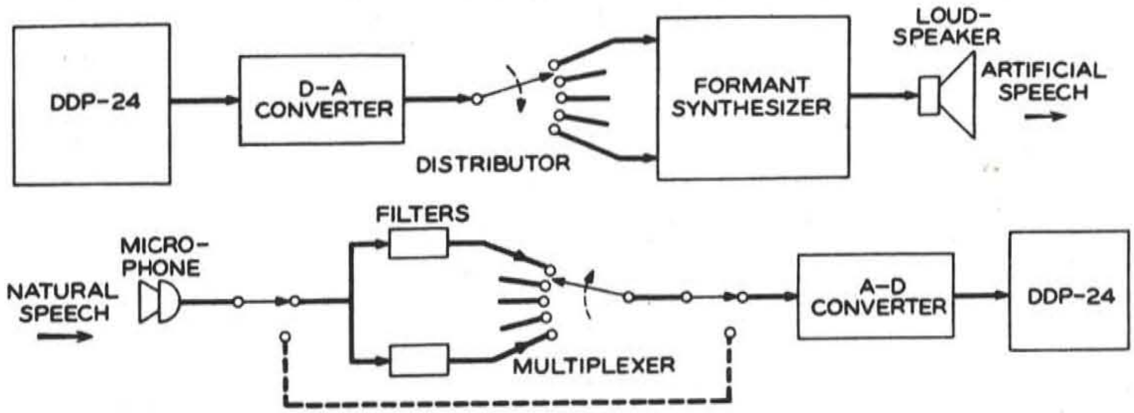

FIG. 1

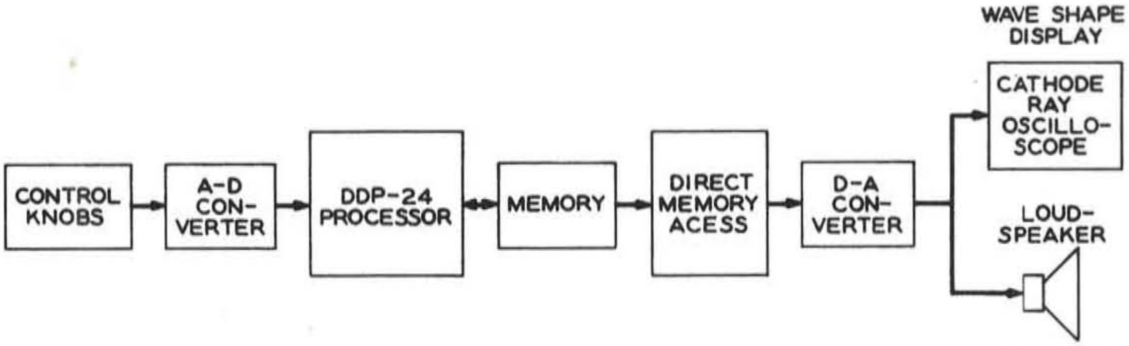

FIG. 2

The third and last requirement for effective on-line operation is to provide the experimenter with good displays of the computed output. Our DDP 24 has provision for presenting its output both visually and acoustically. The time required for the output of data can result in noticeable loss of computing time. For this reason a "direct memory access" channel has been added to our DDP 24. This channel accesses memory independently of the main computer, and data can therefore be transmitted without loss of time while the DDP 24 continues its computation from another part of memory. The general input-output arrangements concerned with data display are shown in Fig. 2.

Two experiments, currently being tried, will serve as examples of on-line speech processing. The first is concerned with the segmentation problem. A speech wave input is inverse filtered to see if boundaries of articulatory nasalization and lateralization could be determined by finding sharp changes in the acoustic properties of the vocal tract. The entire inverse filter process is carried out by computation. The frequencies of two inverse formants are adjustable independently by turning the two knobs, and the experimenter can listen to the inverse filtered speech as well as being able to see its waveshape or spectrum displayed on the screen of a cathode ray tube. The close man-model coupling should help in the rapid understanding of the ways in which inverse filtering can help in the solution of the articulatory segmentation problem. It should also provide better insight into the criteria and decision processes involved in inverse filtering itself - a matter which is as yet not understood well.

The second experiment is concerned with sentence synthesis. A phoneme sequence is first typed into the computer. The computer then calculates the formant tracks for the whole sequence, using stored values for each phoneme which are then modified by stored rules for transitions and inflection. The computed formant tracks are outputed and control the analog formant synthesizer. The knobs are used to vary the position of the longest syllable in the sequence and the value of its duration. Alternatively, the operator will be able to control the syllable duration - fundamental frequency relation of selected syllables. The experimenter hears the synthesized sentences over a loudspeaker and also sees the computed formant tracks on a cathode ray tube display. Valuable understanding of these prosodic features of speech is gained from the ability to observe immediately the effect of the knob-controlled changes.

It is too early as yet to report on the results of these experiments, but there is every indication that on-line computer processing will represent as significant a step forward in experimenting with speech models as did the transition a few years ago from analog models to conventional computer simulation.