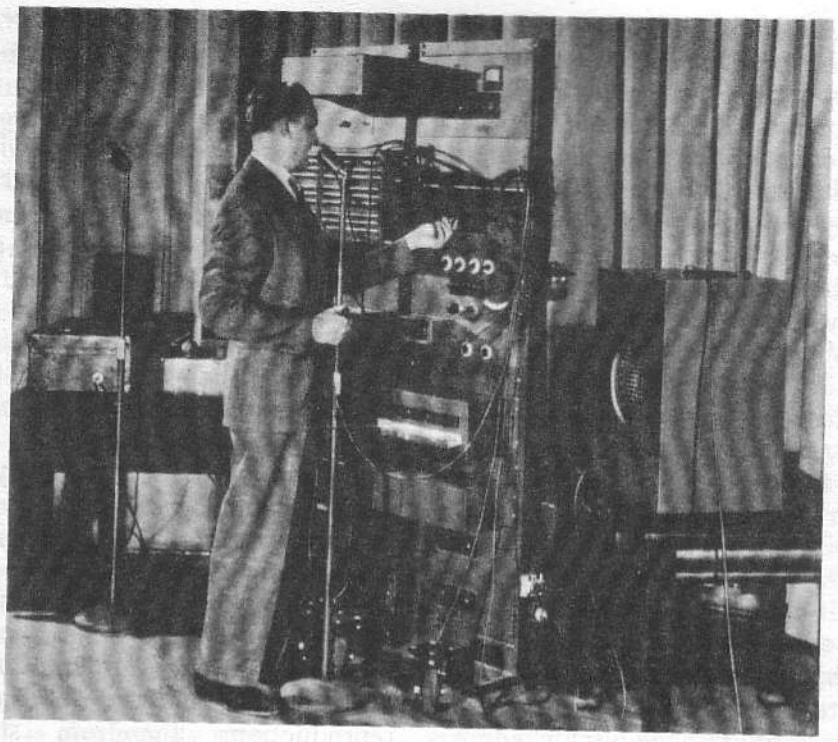

The Vocoder

By HOMER DUDLEY
Circuit Research Department

AT THE World's Fairs in New York and San Francisco great interest was shown in the speech synthesizer in the Bell System exhibits. Known as the Voder, this device creates spoken sounds and combines them into connected speech. Its raw materials are two complex tones, a hiss and a buzz; selection of one or the other and its intensity and tone quality are controlled by an operator through a keyboard.*

The Voder is an offshoot of a more extensive system, first demonstrated† in its experimental stage some three years ago. That system analyzed spoken sounds, and then used the information to control the synthesizing circuit. At the time World's Fair displays were under consideration, so it was naturally perceived that the synthesizer, manually controlled, could be made into a dramatic demonstration. Development was for a while concentrated in that field; as a successful Voder became assured, attention was shifted back to the broader and parent system. Shortly thereafter the system was given the name "Vocoder" because it operates on the principle of deriving voice codes to re-create the speech which it analyzes.

Figure 1 shows the over-all circuit for remaking speech; the analyzer is

*RECORD, *Feb.* 1939, *p.* 170.
†RECORD, *Dec.* 1938, *p.* 98.

at the left and the synthesizer at the right. Electrical speech waves from a microphone are analyzed for pitch by the top channel and for spectrum by a group of channels at the bottom.

In the pitch analysis the fundamental frequency, which for simplicity will be called the pitch, is measured by a circuit containing a frequency-discriminating network for obtaining this frequency in reasonably pure form; a frequency meter for counting, by more or less uniform pulses, the current reversals therein; and a filter for eliminating the actual speech frequencies but retaining a slowly changing current that is a direct measure of the pitch. (Unvoiced sounds, whether in whispering or the unvoiced sounds of normal speech, have insufficient power to operate the frequency meter.) The output current of the pitch channel is then a pitch-defining signal with its current approximately proportional to the pitch of the voiced sound and equal to zero for the unvoiced sounds.

There are ten spectrum-analyzing channels,* the first handling the frequency range 0-250 cycles and the other nine, the bands, 300 cycles wide, extending from 250 cycles to 2950 cycles, a top frequency which is representative of commercial telephone circuits. Each spectrum-analyzing channel contains the proper band filter followed by a rectifier for measuring the power therein and a 25-cycle low-pass filter for retaining the current indicative of this power but eliminating any of the original speech frequencies.

The operation of the analyzer is illustrated in Figure 2 by a group of oscillograms taken in analyzing the sentence "She saw Mary." To insure that the same speech was analyzed in obtaining the various oscillograms, the sentence was recorded on a high-quality magnetic-tape recorder and reproductions therefrom supplied current to the analyzer. The speech-

*A 30-channel vocoder covering the wide range of speech frequencies required for high quality has also been built and is being used as a tool in laboratory investigations.

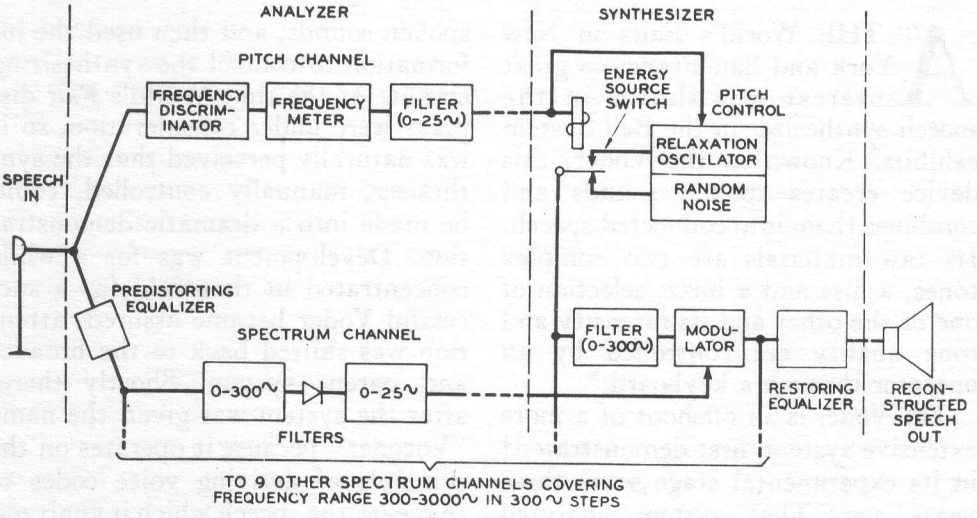

Fig. 1—Simplified schematic of the Vocoder

wave input to the analyzer is shown in the line next to the bottom while the output is shown in the other oscillogram traces; the pitch-defining signal is at the bottom in the figure and the ten spectrum-defining signals in numerical order at the top. For convenient reference the oscillograms are lined up together whereas in the actual circuit the speech-defining signals lag about 17 milliseconds behind the speech-input wave. The inaudible speech-defining output signals contain all the essential speech information as to the input wave, but it is to be noted that they are slow-changing and in this way correspond to lip or tongue motions, as contrasted with the higher audible vibration rates of the rapid-changing speech wave itself. The dropping of the pitch to zero for the unvoiced sounds "sh" and "s" is also readily seen.

Figure 2 gives an idea also as to the synthesizing process. In the analyzer the speech wave is the input and the eleven speech-defining signals are the output; in the synthesizer the eleven speech-defining signals are the input and the speech wave the output.

The steps in speech synthesis are indicated at the right of Figure 1. The relaxation oscillator is the source of the buzz; and the random noise circuit the source of the hiss. The hiss is connected in circuit for unvoiced sounds and for quiet intervals. (In the latter case no sound output from the synthesizer results because there are no currents in the spectrum channels.) When a voiced sound is analyzed a pitch current other than zero is received with the result that the buzz is set for the current pitch by the "pitch control" on the relaxation oscillator; also, the relay marked "energy source switch" operates, switching from the hiss source to the buzz source.

December 1939

The outputs from the spectrum-analyzing channels are fed to the proper synthesizing spectrum controls with the band filters lined up to correspond. The power derived from the energy sources of the synthesizer in these various bands is then passed through modulators under the control of the spectrum-defining currents. The result is that the power output from the synthesizer is sensibly proportional in each filtered band to that measured by the analyzer in the original speech. From the loudspeaker comes, then, speech approximately the same in pitch and in spectrum as the original. This synthetic speech lags the original speech by about 17 milliseconds due to the inherent delay in electrical circuits of the types used.

In the present models of the Vocoder, control switches have been introduced which permit modifications in the operation of the synthesizer. Through the manipulation of these controls interesting effects are produced. Some of the possibilities of the Vocoder were recently demonstrated by the author and his associate, C. W. Vadersen, before the Acoustical Society of America and before the New York Electrical Society. In those presentations Mr. Vadersen supplied by his own voice the incoming speech which was picked up by a microphone as shown in the headpiece; and at the same time he manipulated the controls to produce desired effects. A remote-control switch was also provided through which, for purposes of comparison, the author could switch the microphone directly to the loudspeaker and so let the audience hear how the speech would sound if it had not been modified by the Vocoder.

In these demonstrations comparison is first made between direct speech

and the best re-creation that the apparatus could make. Then by manipulation of dials and switches, speech is modified in various ways. Normal speech becomes a throaty whisper when the hiss is substituted for the buzz. Although the hiss is relatively faint, it is shown to be essential for discrimination as between "church" and "shirts."

Ordinarily the re-created pitch moves up and down with that of the original. If variation is prevented, the re-created speech is a monotone, like a chant. When the relative variation is cut in half, the voice seems flat and dragging; when the swings are twice normal, the voice seems more brilliant; when four times normal it sounds febrile, unnatural. The controls can be reversed so that high becomes low: the tune of a song is then unrecognizable, and speech has some of the lilting characteristics of Scandinavian tongues. Another control fixes the basic value of the re-created pitch; if this is "fluttered" by hand, the voice becomes that of an old person. By appropriate setting of the basic pitch, the voice may be anything from a low bass to a high soprano, and several amusing tricks can be performed. In one of these, the basic pitch is set to maintain a constant ratio of 5 to 4 to the original. This is a "major third" higher and harmonizes with the original. In two-part harmony, the demonstrator then sings a duet with himself. Connecting a spare synthesizer set for a 3 to 4 ratio he then sings one part in a trio, the others being taken by his electrical doubles. Finally, with the basic pitch-control of the apparatus, he becomes a father reprimanding his daughter; then the girl herself, and then becomes the grandfather interceding for the youngster.

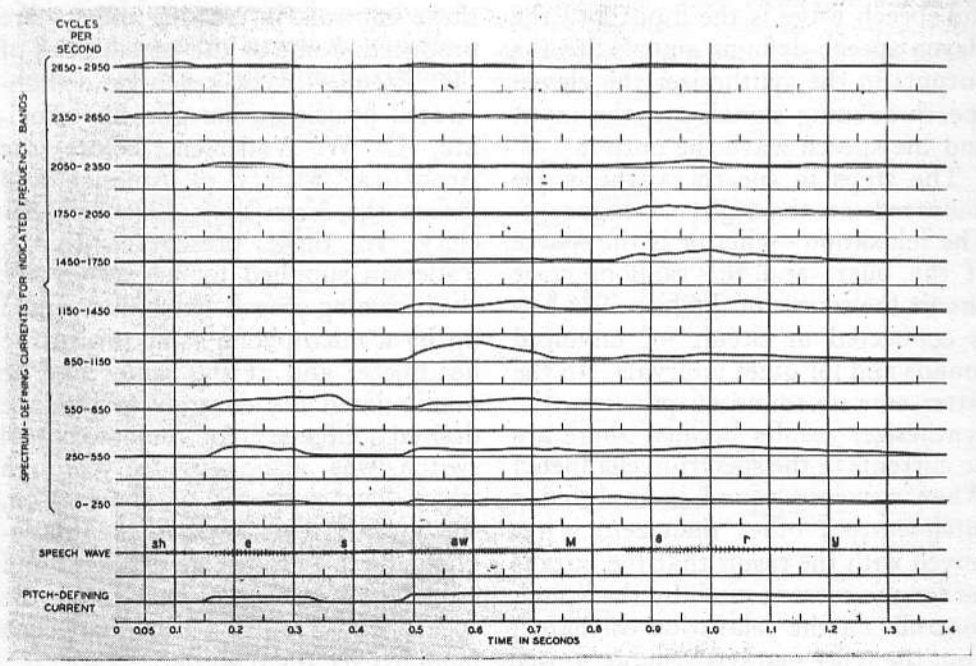

Fig. 2—The original speech wave and an analysis of its components, expressed as the variation of several direct currents

December 1939

For the vocal-cord tones of the original, the Vocoder substitutes the output of a relaxation oscillator. But any sound rich in harmonics can be used: an automobile horn, an airplane roar, an organ. In some demonstrations, the sound, taken from a phonograph record, replaces the buzz input from the oscillator. Keeping careful time with the puffs of a locomotive, the demonstrator can make the locomotive puff intelligibly "We're - start - ing - slow - ly - faster, faster, faster" as the puffs come closer together. Or a church bell may say "Stop - stop - stop - don't - do - that." A particularly striking effect is that of singing with an organ to supply the tones. Although the words may be spoken, the demonstrator usually sings them to hold the rhythm. It makes no difference whether his voice is melodious or not; the tonal quality comes only from the musical source.

These tricks and others have suggested uses for the Vocoder in radio and sound pictures. It appears to have possibilities as a tool in the investigation of speech, since by its numerous controls important variables in speech can be isolated for study. The engineering possibilities which may grow out of the application of the principles employed in this device are hard to predict at the present time. The speech-defining currents, however, do have features of simplicity and inaudibility which may open the way to new types of privacy systems or to a reduction in the range required for the transmission of intelligible telephonic speech.