Digital Computer Synthesizes Human Speech

Two basic approaches to the production of synthetic speech will be presented by John L. Kelly of the Visual and Acoustics Research Laboratory to a seminar in Stockholm next month. Research in this field represents an attempt to understand the basic phenomenon of speech, as well as understanding the necessary elements in the transmission of speech. Such understanding is essential to an efficient approach to telephone transmission in the future.

The two approaches to be discussed at Stockholm aim at the same goal: the generation of speech from an input consisting of *names* of the elementary sounds or phonemes, plus a minimum amount of information on timing, stress, and inflection. The first approach involves a "terminal analog," or a machine such as a vocoder, whose inputs are acoustical parameters, such as pitch, buzz intensity, and formant frequencies. The other approach uses a "vocal tract analog," whose control signals represent articulatory parameters such as the shape of the vocal tract, nasal coupling and tongue position. In either case a set of rules must be worked out to generate control signals from phonetic information.

To compare the virtues of the two approaches to the production of synthetic speech, Dr. Kelly produced computer programs which simulate both the "speaking machines" proposed. He then recorded the output of the computer on audio tape. Tapes of both types of machines will be played at the seminar.

The opposite page is a recording of a tape produced using the first mentioned method—the terminal analog. This machine, proposed jointly by Mr. Kelly and Louis J. Gerstman, is of the tandem resonant type, with several novel principles used. The computer used to simulate the "speaking machine" was programmed to accept in sequence the names, on punched cards, of the phonetic speech sounds which make up an English sentence. The computer then processed this information the way an actual speaking machine would, and produced an output like the output of the speaking machine.

The program had two parts. One simulated the speaking machine; the other consisted of rules, derived from previous research, for combining the individual speech sounds into connected speech and producing control signals for driving the speaking machine. Nine control signals corresponding to voice pitch, buzz intensity, and hiss intensity, plus the center frequencies and bandwidths of three speech formants were continuously generated.

The speech of the simulated talking machine came out of the computer on digital magnetic tape, and was then converted to a variable magnetic sound track suitable for playing on an ordinary tape recorder.

On the demonstration tape recorded here, the computer "says" simple sentences in a measured monotone voice. Then more natural inflection and phrasing is inserted. This was obtained by specifying the changes in pitch and timing on each punched card.

When the pitch of the sounds is varied, the computer can also be made to sing, as witnessed by the recording of "Bicycle Built for Two"; also a few lines of the "To Be Or Not To Be" soliloquy from *Hamlet* are included.

The samples presented are early results of a research project by Kelly and Gerstman to obtain a better understanding of the nature of speech. Ultimately this knowledge may be useful in devising new ways of transmitting speech efficiently over communication systems. For example, a person may, in the future, be able to sit at a keyboard and by typing, cause a talking machine thousands of miles away to speak for him.

There is also the possibility that talking machines, like the one simulated in the computer, could be built for use by people who are unable to speak. By typing the phonetic symbols on a keyboard they could direct a talking machine to speak for them.

Also, in the future, a blind person may be able to have a speaking machine read to him from books which have been previously encoded on a punched tape.

33 1/3 rpm microgroove

TEAR OUT AND PLAY

Produced by Bell Telephone Laboratories

To prevent this paper record from skidding, tape corners to a standard record or the turntable.

BAND 1 The computer speaking **BAND 2** The computer reciting a soliloquy from Hamlet **BAND 3** The computer singing

Samples of speech generated by an electronic digital computer which has been programmed to read phonetic symbols and utter speech-like sounds.

NOTES→

Bell Telephone Laboratories Presents

SYNTHESIZED SPEECH

The samples of speech on this recording were produced by an electronic digital computer. They are a by-product of a research project at Bell Telephone Laboratories to obtain a better understanding of the nature of speech. Ultimately, this knowledge may be helpful in devising new ways of transmitting speech over communications systems.

A Machine That Talks

There are many possible kinds of speech synthesizers or "talking" machines. To save the expense and time of building, testing, and modifying them, John L. Kelly, Jr. and Louis J. Gerstman of the Visual and Acoustics Research Department at Bell Laboratories use a high-speed, general purpose computer to simulate them. The computer is instructed to accept certain information on punched cards, to "operate" on this information similarly to the way an actual talking machine does, and to produce an "output" analogous to the output of the talking machine. By changing the computer program it is comparatively easy to modify the characteristics of the talking machine.

The particular machine which was simulated in the computer to produce the speech on the recording is known technically as a "tandem resonant synthesizer." Usually, this type of machine is operated by continuously feeding into it a set of nine signals corresponding to voice pitch, voice loudness, tongue position, and other speech variables. When every instant of sound is specified, the machine produces sounds that are amazingly like human speech.

A Phonetic Input

Doctors Kelly and Gerstman have contributed a significant advance in the art of speech synthesis by devising a computer program which permits them to feed into the computer, on punched cards, the names of speech sounds. Since the standard phonetic symbols representing speech sounds are not included on the keyboard of an ordinary card-punching machine, Kelly and Gerstman devised a new phonetic code using the letters of the alphabet. At present it consists of 22 consonants and 12 vowels:

Consonants: P,B,T,D,K,G,M,N,NG (si**ng**), F,V,S,Z,SH(**sh**e), ZH (a**z**ure), H,W,R,L,Y,TH(**th**in), DH (**th**en).

Vowels: EE(b**ee**), I(**i**ll), AY (r**a**te), E (**e**nd), AE(**a**dd), AH(**ah**), AW(j**aw**), O(g**o**), OO(f**oo**t), UU(f**oo**d), UH (**u**p), ER(h**er**).

Each speech sound is specified on a separate punched card. When a sequence of cards is fed into the computer, the stored computer program "operates" on this information to produce the nine control signals for driving the simulated talking machine.

For example: if a sequence of cards H,EE,S,AW,DH,UH,K,AE,T is put into the computer, the talking machine will say "He saw the cat" in a measured monotone voice. To obtain natural intonation and phrasing it is necessary to specify on each card (in addition to the speech sound) both the pitch of the sound and timing information.

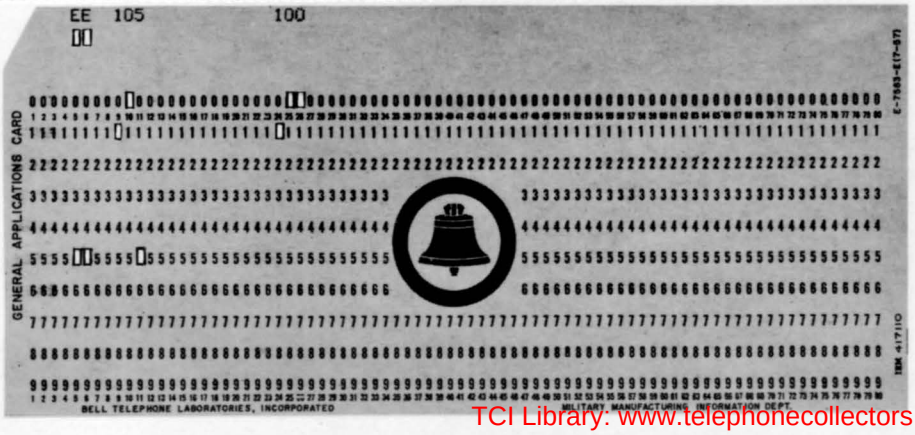
A Speech-like Output

The "speech" of the simulated talking machine comes out of the computer in the form of tiny magnetized spots on half-inch magnetic tape. This tape is then fed to another machine which converts the digital information to a variable magnetic sound track suitable for playing on an ordinary tape recorder playback.

On This Recording

The samples of speech on this recording illustrate the present state-of-the-art of speech synthesis.

33-1/3 rpm microgroove



SPEECH SOUNDS are specified on standard punched cards in the new synthesized speech program devised by Kelly and Gerstman. This particular card instructs the computer to produce the sound EE (as in beat) at a pitch of 105 cps for 100 milliseconds.