Speech synthesis

Design case history: Speak & Spell learns to talk

Designers of an electronic learning aid mastered advanced speech technology and meticulously 'prepared' the consumer

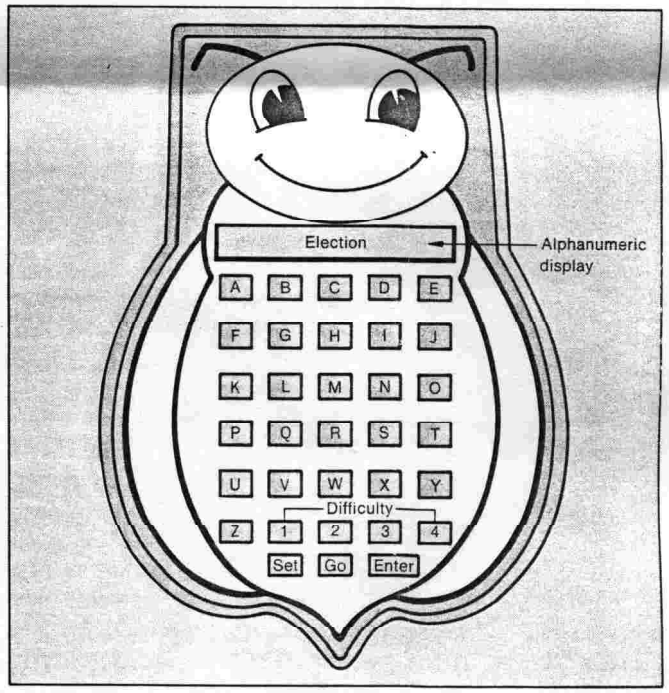
Four Texas Instruments engineers, each a specialist in a different field of electronics, got together one day in 1976 at the company's headquarters in Dallas for a brainstorming session. Their aim was simple enough: to develop a new electronic learning aid for children.

Earlier that year TI had introduced the Little Professor, an electronic aid that taught basic arithmetic, and its reception by the public was so encouraging that the company was looking for an encore.

As the four engineers riffled through ideas, they were drawn to one: an electronic spelling bee. As outlined by the computer specialist on their panel, Paul Breedlove, it would be a major innovation over the Little Professor concept, and technologically it would be a more complex nut to crack than math.

Visual displays might work for most nouns, but how do you enable a child to visualize verbs and such abstract words as "courage" or "jealous" without giving away the correct spelling? In school, children usually learn spelling with a teacher who calls out the word and has the child spell and pronounce it. The new TI aid would need a voice, the panel agreed. But there was the rub: high-performance techniques then in use called for many arithmetic operations and large computer memories. Such a product would not be portable. On top of that, it would cost thousands of dollars.

Eighteen months later, after the idea had filtered through many technical and marketing channels at TI, the Speak & Spell learning aid emerged. It met the goal of a hand-held, low-cost spelling product with speech output.



[1] The original model electronic learning aid for teaching spelling to children was the Spelling Bee. The concept was an outgrowth of the development of the Little Professor, Texas Instruments' first learning aid.

A high-risk project

Besides Mr. Breedlove, the others on that brainstorming panel were Gene Frantz, a project engineer, who was put in charge of the spelling bee project; Larry Brantingham, whose background was in integrated-circuit design; and Richard Wiggins, a mathematician, who took over the work of specifying the speech synthesis algorithm.

As proposed [Fig. 1], the product had many uncertainties. How could the speech be generated? Would the words be intelligible? Would the product justify the development effort? The four engineers took their idea first to the corporate Objectives, Strategies, and Tactics Committee, the body at TI that determines how funds are invested in new concepts [see the special issue on productivity, *Spectrum*, October 1978, pp. 78–80]. The engineers were encouraged to seek funding through a less formal routine—the company's Idea program. This program funds the initial stages of high-risk projects that would otherwise find it impossible to compete for company funds. Once a program becomes better defined, further money may be appropriated by the Objectives, Strategies, and Tactics Committee.

The Spelling Bee program was granted $25 000 for three months to demonstrate its technical feasibility. At a series of meetings in early December 1976, members of the corporate research laboratories and the product and component design departments of the Consumer Products Group agreed that to meet the goal of portability and low cost, a technological breakthrough was needed.

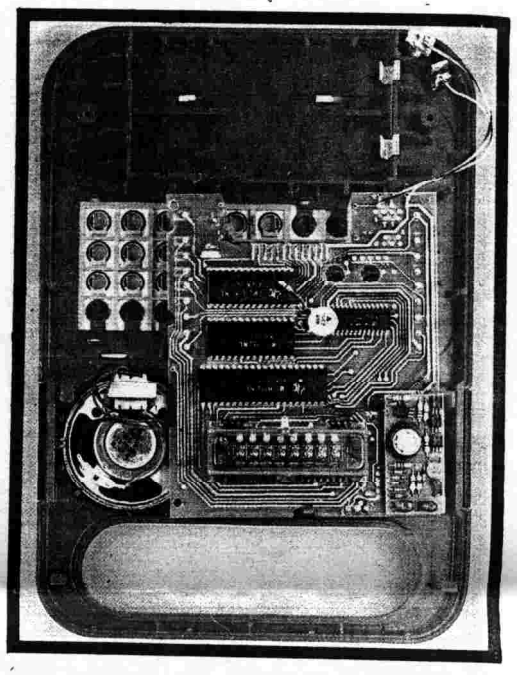
Alternative synthesis techniques considered

A number of speech-synthesis techniques were considered, beginning with a synthesizer that operated from stored phonemic descriptions, with a set of rules for adding pitch, inflection, stress, and timing. This approach was abandoned because the synthesizer could not be a simple digital circuit and because of concerns about the recognition of single words out of context. Some form of an analysis synthesis technique was preferred for three reasons. First, this approach offered superior voice quality,

Gene A. Frantz and **Richard H. Wiggins**
Texas Instruments Inc.

0018-9235/82/0300-0045$00.75©1982 IEEE

[2] As finally produced, the Speak & Spell learning aid took on a more serene form. Packaging problems were minimized by the development of the synthesizer chip. This IC operates in conjunction with a four-bit microprocessor and two 128-kilobit ROMs.



which was of primary concern. Second, although the data rates would not be as low as with the phonemic synthesizer, they would be sufficiently low for a moderate vocabulary for this product. Third, the necessary data processing required for the speech analysis procedure could be developed at minimum cost.

Of the two most likely analysis techniques—formant synthesis and synthesis using linear prediction (LP)—the latter was chosen because it achieved higher speech quality at low data rates.

The major problem in implementing a synthesizer with linear-predictive coding was the many arithmetic operations to be performed in the digital filter. A two-multiplier lattice filter was chosen because computations could be done in fixed-point arithmetic and also because reflection coefficients could be used directly. This was considered important inasmuch as reflection coefficients are well suited for coding. At a 10-kilohertz sample rate, however, the process would require 200 000 multiplication operations per second and a similar number of additions.

Besides the digital filter, the LP synthesizer consisted of a decoder, a parameter smoother, an excitation generator, and a digital-to-analog converter. Each of these functions had unique performance requirements that were met with a minimum chip size.

After three months a computer simulation of the basic processing architecture and the chip area was completed. The simulation showed that integrating a complete 1200-bit-per-second speech synthesizer on a single chip was now possible.

Technology is sold to management

The next step was to sell the technology development program to management. The project director requested and received additional funds from TI's Objectives, Strategies, and Tactics Committee. The company was aware of a few market truisms. The Speak & Spell (TI tradename) learning aid was planned, of course, as a large-volume consumer product. Large-volume markets can create the motivation for the development of high-technology custom circuits. The completed IC, in turn, drives down the unit cost of the product, and the lower unit cost opens the way to large market demand. This closed feedback system usually results in rapid development and application of high technology. To a large degree, the Speak & Spell learning aid had all the necessary elements of such a feedback system.

Nevertheless there was some uncertainty over whether the speech requirements would be met. The development program was funded so that money for a new phase was based on the successful completion of the previous one. Obviously, as each milestone was reached, confidence increased, and funding became easier to obtain.

Parallel to the efforts of selling solid-state synthetic speech to management was laying the groundwork for marketing the new talking learning aid. The advantages of verbal interaction between the product and the student were expected to be obvious to the consumer, but this was soon proved wrong for several reasons. The average person had never listened to solid-state speech before. Exposure to talking machines was limited to movies like *2001: A Space Odyssey,* in which the talking machine, Hal, was cast as the bad guy. Thus, TI learned, many consumers associated the characteristics of synthetic speech with a dull monotone. What's more, the public was used to toys that were made to speak by a tug on a string or the operation of an analog cassette tape.

The company's marketing gurus tested consumers to determine their acceptability of TI's planned product. Subjects were selected randomly and grouped into six categories: one group of fathers, three of mothers, and two of teachers. The children these adults came into contact with were ages 7 through 10. TI introduced the test groups to its product with the aid of posters and voice samples played on a cassette tape recorder. The results were disappointing. Major drawbacks found by parents and teachers were as follows:

- A 250-word capacity was too limited.
- The words were pronounced with an unacceptable accent and in a cold and computerlike manner.
- The product would be unreliable, since the talking toys then sold were, in general, broken by children.
- It would be just another noisemaker.
- Children would quickly become bored and lose interest in the product.

At TI's laboratories, meanwhile, researchers were struggling to develop a read-only memory that could store at least 300 words without increasing the product's cost unduly. Engineers designed the largest ROMs in the industry at the time: 131 072 bits of stored information. This capacity, however, could store

only 150 words in each ROM. Two ROMs were tried to give the product a vocabulary of 300 words and phrases, but consumers indicated that this still was not enough. TI engineers then designed speech ROM plug-in modules, which could be sold separately to add words to the product.

The concern that a talking toy would not hold up under use was based on the fact that such toys were mechanical and were more abused than used by small children. The Speak & Spell learning aid, however, had no moving parts. This advantage was heavily stressed in TI's advertisement campaign.

To counter objections that the learning aid would quickly degenerate into a monotonous noisemaker, with the same

How speech was compressed in Speak & Spell

Besides the linear-prediction analysis procedure in the speech compression for the Speak & Spell learning aid, each type of speech information was coded differently. In doing so, the data rate was minimized to less than 1200 bits per second without degradation of the speech quality.

The different speech characteristics and the bits per frame needed to describe each were as follows:

Type	Number of bits per frame	How determined	Data included
Voiced	49	E≠0,15;P≠0;R=0	E,R,P,K1-K10
Unvoiced	28	E≠0,15;P=0;R=0	E,R,P,K1-4
Repeated frame	10	E≠0,15;R=1	E,R,P
Zero energy	4	E=0	E
End of word	4	E=15	E

Coding for the different speech characteristics was based on several assumptions. First, speech sounds are of two types: voiced and unvoiced. Generally vowels, which are periodic in nature, can be considered voiced sounds, and nonperiodic sounds, such as many consonants, can be considered unvoiced sounds. To describe a speech frame (25 milliseconds of speech), 49 bits are needed for a voiced sound and 28 for an unvoiced sound.

A second assumption was this: many voiced and unvoiced sounds do not significantly change in character for long periods of time. A voiced sound may not change for several frames, perhaps for 100 ms. When this occurs, it is not necessary to send the same 49 bits, in the case of a voiced sound, to describe each frame. For a 100-ms voiced sound, the first frame can be described by a 49-bit frame and the following three repeat frames tell the synthesizer to use the previous vocal-tract parameters in the present frame of speech. Since a repeat frame is 10 bits long, a saving of 39 bits per frame was realized for a voiced frame and 18 bits per frame for an unvoiced frame.

Zero-energy frames are used to describe periods of silence in speech. At such times the pitch and vocal-tract parameters are not necessary. Therefore only 4 bits of energy are used. Finally, an indication of when a word ends is treated as a special case of the energy parameter; this indicator tells the synthesizer to stop speaking.

As an example, if the energy value (E) is either zero (0000) or 15 (1111), no more data are needed to describe the frame, since it is either a zero-energy frame or an indication of the end of a word. If the energy is nonzero, then the repeat bit is needed (the fifth bit) to determine what additional information is necessary to describe the speech for the frame. If the repeat bit (R) is equal to 1, then only the pitch information (next 5 bits) is needed, and the previous set of reflection coefficients (K values) are used to describe the vocal tract model of the present frame. If R is equal to zero, then the pitch information must be used to determine how many reflection coefficients are required. If the pitch information is equal to zero, it is an unvoiced frame, and only the first four reflection coefficients (K1–K4) are used. If the pitch information is not equal to zero, then it is a voiced frame, and all 10 reflection coefficients are needed.

The accompanying coded data for the word HELP illustrates how the data length of each frame is determined.

—*G.A.F. and R.H.W.*

	E	P	R	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
	0000												
HEL	0100	00000	0	10011	01110	1001	0111						
	0111	00000	1										
	1101	10010	0	10000	10100	1000	0110	0111	1000	1010	100	101	010
	1101	10011	1										
	1110	10011	1										
	1101	10100	0	01101	01111	1010	1010	1001	0111	1000	100	101	101
	1101	10100	0	01110	01011	1000	1100	1101	1000	0100	100	011	101
	1101	10011	0	10001	01010	0110	1001	1111	1011	0101	010	000	110
	1011	10010	1										
	1010	10010	0	01101	00111	1000	1100	1111	0111	0010	001	010	110
	1001	10001	1										
	1001	01110	1										
	1000	01101	1										
	0010	01110	0	00101	00101	1101	1001	1110	0101	0111	001	011	011
	0000												
	0000												
	0000												
P	0111	00000	0	10100	01011	1011	1000						
	0111	00000	0	10001	01011	1011	0110						
	0101	00000	1										
	0011	00000	0	10011	00111	1010	0110						
	0010	00000	0	10010	00101	1011	0101						
	0000												
	1111												

phrases repeated over and over, the company selected multiple phrases for the responses, some randomized and some sequential. The result was a product that seemed to communicate in a human manner.

The one constraint placed on the selection of phrases concerned the error messages. If the child's spelling was wrong, the Speak & Spell learning aid used only one phrase the first time: "Wrong, try again." If the child's next attempt was also wrong, the machine chided: "That is incorrect. The correct spelling of . . . is. . . ." Some engineers thought it would be fun for the child to receive raspberries or a catcall or some funny comment for a wrong spelling. But responses like these were rejected by the developers because, while more exciting for the child, they would tend to "reward" incorrect spellings.

The next issue that had to be resolved was the matter of word pronunciation. Picking the "correct" American dialect was not easy. However, for the product to speak correctly the designers eliminated some early suggestions for a character-type voice. Instead, the final choice was a voice compatible with the analysis system and with correct pronunciation.

Finally, that children would become bored was not only a concern of the test group but also of the design team. Some test subjects may have perceived the product as similar to a tape recorder. Others may have assumed it would be a mere list giver, with no interaction with the child. To alleviate these fears, the engineers developed several word games so a child could engage in a variety of activities, all designed to build word skills.

The Speak & Spell learning aid was introduced in June 1968 in Chicago at the Consumer Electronics show [Fig. 2]. It was well received from the start.

In retrospect, the fears associated with speech synthesis and public acceptance of the product were clearly unfounded. As

How a speech synthesizer works

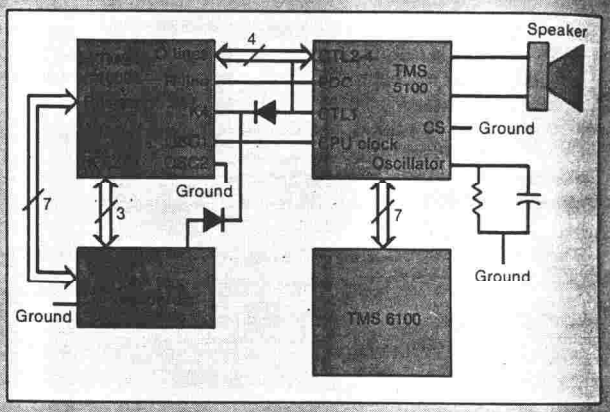
The TMS 5100 speech synthesizer chip [below] uses a tenth-order linear-predictive coding model (LPC-10). The model is controlled by 12 parameters: pitch, energy, and 10 reflection coefficients. These parameters are stored in an external read-only memory and are used to update the synthesizer every frame, or approximately every 20 to 25 milliseconds.

The output of the model drives a digital-to-analog converter, which in turn directly drives a mechanical speaker. The interface bus is controlled by a chip-select line and a synchronization signal. A master clock can be driven by either an external resistor and capacitor or by a ceramic resonator. Two clock outputs drive the microprocessor and the speech ROM. A push-pull amplifier drives a 100-ohm center-tapped speaker.

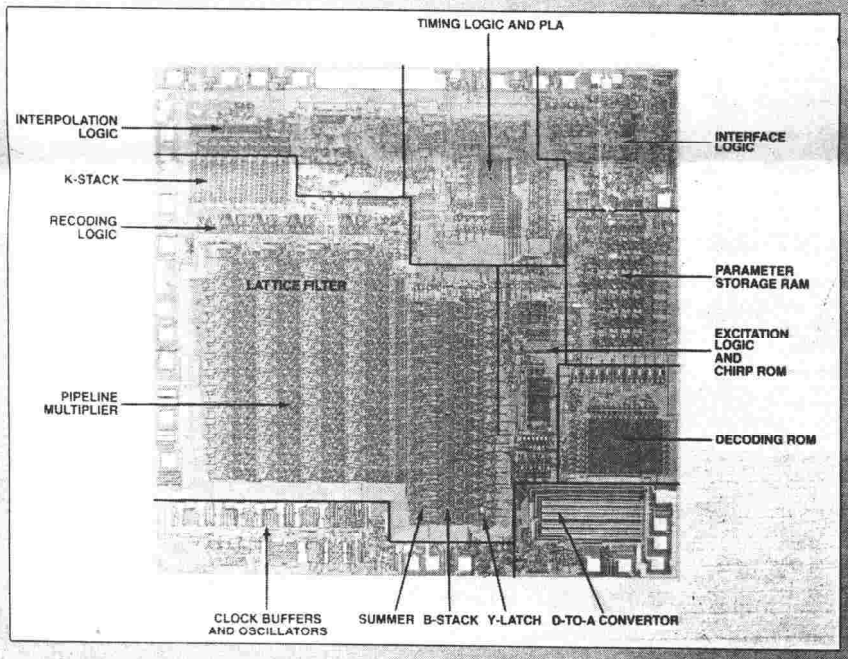
In the Speak & Spell learning aid, the TMS 5100 is designed to work with a custom microprocessor, the TMC 0270. The TMS 5100 was also designed to work with the TMS 1000 4-bit microprocessor family. Several different configurations can be used. One is shown at right. Here five output lines are dedicated to the synthesizer, along with one of the input lines. The chip-select line is tied to ground.

A possible second configuration could allow the output lines to be shared with other tasks if the six output lines were used to send instructions to the synthesizer. One of the lines would be dedicated to the TMS 5100. The five others could then be used for other functions. As in the first configuration, one input line of the TMS 1000 would be tied to one of the communications lines to accept data from the synthesizer.

—*G.A.F. and R.H.W.*



One configuration for the TMS 5100 synthesizer chip is to use it with the TMS 1000 4-bit microprocessor and the TMS 6100 128-kilobit ROM. This configuration uses minimal components including an RC oscillator, a keyboard, a couple of diodes, and the speaker.



The bulk of the synthesizer chip is taken up by the pipeline multiplier, which is a part of the lattice filter. Included on the same chip is a ROM that stores the decoding information and a RAM for storing the voice parameters. Dimensions are 214 mils wide by 210 mils deep.

Solid-state speech synthesis expanding

When the Speak & Spell learning aid was introduced in 1978, one other solid-state speech-synthesis product was already on the market. Telesensory Systems Inc. of Palo Alto, Calif., had been selling its Speech Plus calculator since 1975. Unlike Speak & Spell, which was marketed to the general consumer to teach children to spell, the Speech Plus calculator had a narrower function: to teach arithmetic to blind people. Its development, like the Texas Instruments Speak & Spell, was funded internally.

A paper delivered by Telesensory engineers at the 1976 National Computer Conference shows similarities between the design efforts of Telesensory and TI. Speech Plus had the following general specifications:

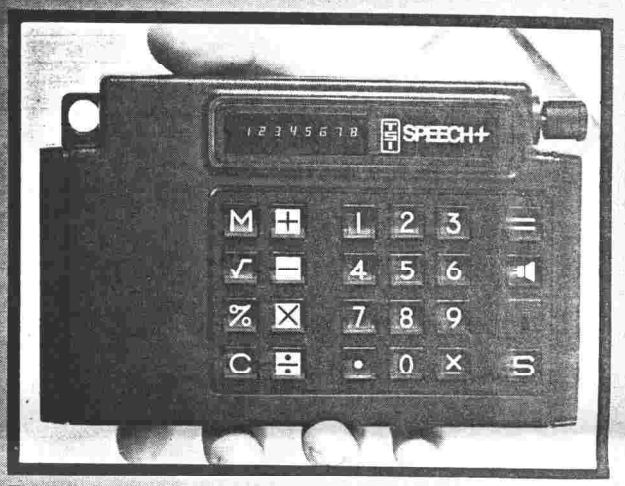
- The calculator had to be hand-held and battery operated.
- Spoken speech had to be generated for every key that was depressed and on command for the display.
- The product had to be cost-effective and reliable.

A custom LSI speech-synthesizer chip was designed and fabricated in MOS technology. The chip reduced a prototype synthesizer, comprised of 60 ICs, to a single microcontroller chip and a ROM.

The controller chip implements a proprietary algorithm for speech synthesis. Control information and data are accessed from the ROM by a separate calculator circuit. The speech output is a 24-word vocabulary, as opposed to TI's 320 words and phrases. The Telesensory device uses a less sophisticated waveform-analysis technique that was suitable at the time for its limited vocabulary.

Since then Telesensory has announced a speech-synthesizer chip that uses linear-predictive coding. When sold as part of a printed circuit that also has ROMs, the IC typically produces more than 100 words or utterances and operates at a 2200-bit-per-second rate. The chip is being sold in the original-equipment-manufacturer market, where companies like Texas Instruments, the Votrax Division of Federal Screw Works, Toshiba, Hitachi, and National Semiconductor already have versions of their own. As a result, manufacturers are applying speech synthesis to many other uses besides educational. Applications include personal computers and "white" goods like microwave ovens, refrigerators, and washing machines, as well as industrial uses in computerized inventory and distribution systems and manufacturing systems.

—*Nicolas Mokhoff*



The portable Speech Plus talking calculator was developed for the blind in 1975. A custom microcontroller and a 16-kilobit ROM synthesized 24 words.

with any new venture, some advice had to be ignored, risks had to be taken, and compromises had to be made. Despite that, the product changed only slightly from its original concept. And from the beginning of the program until the introduction of the product 18 months later, all schedules were met.

To probe further

IEEE Spectrum has published numerous articles dealing with various aspects of speech synthesis. A few major articles in recent issues were:

D.C. Songco *et al.*, "How computers talk to the blind," May 1980, pp. 34–38.

B.A. Sherwood, "The computer speaks," August 1979, pp. 18–25.

J.L. Flanagan, "Synthetic voices for computers," October 1970, pp. 22–45.

Articles on speech synthesis are published regularly in the *IEEE Transactions on Acoustics, Speech, and Signal Processing*, published six times a year.

A special issue of the *Proceedings of the IEEE*, April 1976, was devoted to man-machine communications. Speech-synthesis techniques were covered extensively in, among others, the article by J. Allen, "Synthesis of Speech from Unrestricted Text," pp. 433–42.

Two IEEE Press books have covered speech analysis and speech processing. *Speech Analysis* was edited by R.W. Schafer and J.D. Marke and includes reprints of papers on speech-analysis methods and a compilation of analysis/synthesis systems. The book was published in 1979. It is available clothbound (Order No. PC01123) at $27.70 to IEEE members, $36.95 to nonmembers, and paperbound (Order No. PP01131) at $18.45 to members only. Use the order form in this issue. The second book, *Automatic Speech & Speaker Recognition*, edited by N.R. Dixon and T.B. Martin, contains selected papers (38 in all) on the antithesis of synthesis—speech recognition. It is available clothbound (Order No. PC01149) at $24.70 to IEEE members, $32.95 to nonmembers, and paperbound (Order No. PP01156) at $16.45 to members only.

About the authors

Gene Frantz (M) is manager of the speech technology branch for Texas Instruments' Consumer Products Group. He joined the company in 1974 and since 1976 has been responsible for developing educational products such as the Little Professor and the Speak & Spell learning aid. He was elected a member of TI's technical staff in 1979. Mr. Frantz received a B.S.E.E. from the University of Central Florida in 1971 and a M.S.E.E. from Southern Methodist University in 1977.

Richard H. Wiggins (M) is manager of technology and systems engineering at the company's Speech Technology Center. Prior to this he was a senior member of technical staff at TI's Central Research Laboratories where he developed systems for the analysis and synthesis of speech. Dr. Wiggins holds several patents in the speech-processing area, including some on the LPC-10 speech-synthesis ICs used in the Speak & Spell learning aid. He received B.S. and M.A. degrees in mathematics from Louisiana State University, Baton Rouge, and from the American University, Washington, D.C., respectively. He also earned an M.S. and a Ph.D. in applied mathematics from Harvard University, Cambridge, Mass. ◆