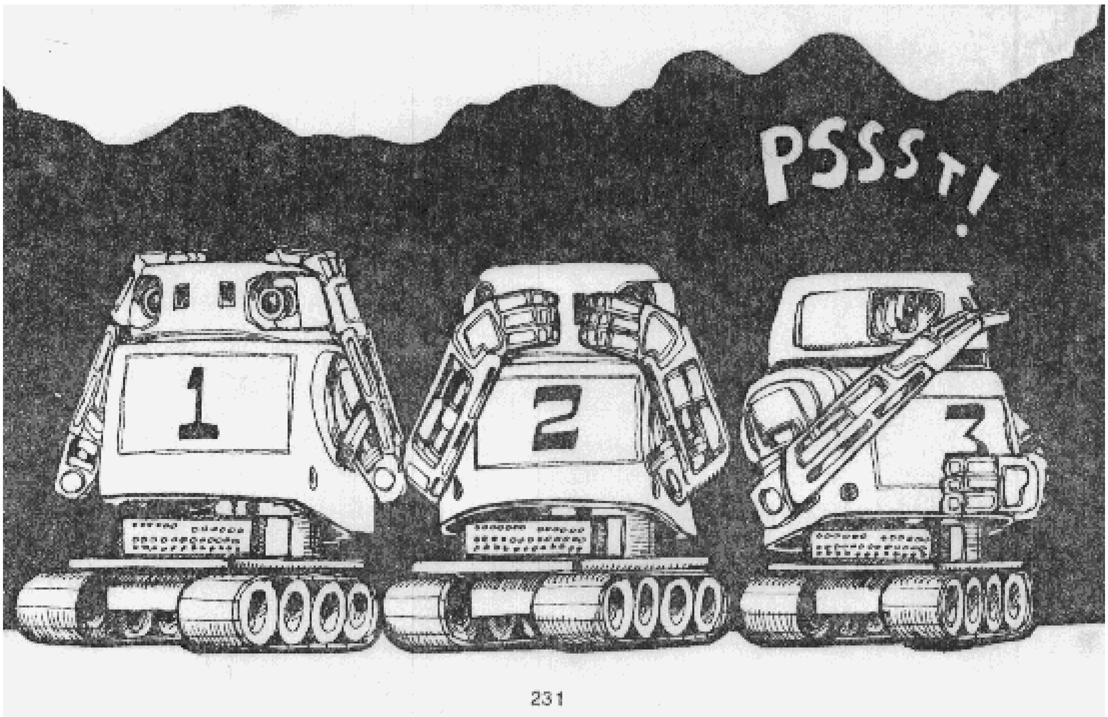
sp synth Altair/Imsai

The Time Has Come to Talk

"The time has come, "the Walrus said, "To talk of many things: of shoes - and ships - and sealing wax - of cabbages - and kings - And why the sea is boiling hot - And whether pigs have wings."
- Lewis Carroll, 1871, in Through the Lookinglass.

Dropbox



Wirt Atmar, Ai Cybernetic Systems, PO Box 4691, University Park NM 88003
published in ????? 19??

The extent to which art and literature, particularly science fiction, affect the future course of civilization remaims a persistent and perplexing question. Must a dream, by necessity, occur decades before its realization? Or does the presence of the dream itself generate its own reality? Mankind's trip to the Moon in 1969 was the dream dreamt by Cyrano de Bergerac and Johannes Kepler 300 years prior to its enactment.

And now, we, nurtured by the thousand different dreams of the future as portryed in novels and movies, all expect computers to be able to talk in the near future. Whether we see the computer becoming the benign and obedient servant of man or wildly out of control, we all tend to see the computer becoming more anthropomorphic, more humanlike in behavior and form.

In science fiction two great dreams of the future predominate. One is the seemingly inevitable first contact with intelligent beings of an extra terrestial origin. The second is the construction, by our own hands, of an alternate embodiment of intelligence in machine form. The first dream may well not occur within the lifetime of our civilization; the second would seem to be almost guaranteed within the next 100 years.

The addition of speech to the computer's behavioral repertoire makes the computer no more intelligent nor aware than it was before. It remains a simple machine. But it undeniably takes on a human characteristic that it never possessed before. An observer finds it impossible not to personify the machine with an identity and a distinct personality. While the addition of speech is only a minor step toward achievement of a truly self-organizing, artifically intelligent machine, it is a psychologically important one. The computer, once it speaks, seems to be intelligent. But again, the dream of machine produced speech is much older than its reality. The ancient Greco-Roman civilization was fascinated with the idea of deus ex machina. Stone gods were often hollowed to allow a priest to speak from within, a practice that persisted well into the Christian era.

The first known practical realization of machine generated speech was accomplished in 1791 by a most ingenious engineer, Wolfgang von Kempelen, of the Hungarian government. Von Kempelen's machine was based on a surprisingly detailed understanding of the mechanisms of human speech production, but he was not taken seriously by his peers due to a previous well publicized deception in which he built a nearly unbeatable chess playing automaton. The "automaton" was unfortunately later discovered to actually conceal a legless Polish army ex-commander who was a master chess player.

By 1820, a machine was constructed which could carry on a normal conversation when operated by an exceptionally skilled person. Built by Joseph Faber, a Viennese professor, the machine was demonstrated in London where it sang "God Save the Queen." Both the Von Kempelen and Faber machines were mechanical analogs of the human vocal tract. A bellows was provided to simulate the action of lungs: reeds were used to simulate the vocal cords, and variable resonant cavities served to simulate the mouth and nasal passages.

The basic method, modelling the human vocal tract, remains to this time the only practical method of actually synthesizing speech. In the 20th century, such modelling is done electronically. The approach was first put in electrical analog form by Bell Laboratories in the late 193Os. The Bell Telephone VODER (Voice Operation DEmonstratoR) was initially shown at the 1939 New York's World Fair where it drew large crowds and considerable attention. The VODER consisted of a buzz source (similar to human vocal cords or mechanical synthesizers), a hiss source to simulate the rush of aspirated air, and a series of frequency filters to imitate the three, four, five or six preferred frequencies (called formant frequencies) passed by the resonant cavities formed by the mouth, tongue and nose. The original VODER was played by highly trained operators using a keyboard, wrist switches, and pedals at an organ-like console. Twenty four telephone operators were trained six hours a day over a 12 month period for the 1939 World's Fair. The VODER itself was a full rack in height.

With the advent of digital computers, however, the synthesis of speech has been made much easier. All the information necessary to repeatedly and reliably generate any one speech sound (a "phoneme") can now be programmed into the machine. Through the proper connection of phonemes, a digital computer could be made to say words and sentences.

General American English, the dialect spoken in the midwest and southwestern parts of the United States, contains 38 distinct phonemes. These speech sounds can be divided into the following classes:

> Pure vowels: produced by a constant excitation of the larynx and the mouth held in a steady position; eg:"e"
>
> Diphthongs:a transition from one pure vowel to another, thus are not always considered as separate phonemes; "i", "u"

Fricatives:consonants produced by a rush of aspirated air through the vocal passages:"f", "s"
Plosives: explosive bursts of air: "p", "k", "t"
Semi-vowels:"w", "y"
Laterals:"l" , "r"
Nasals: "n", "m"

Figure 1:The serial analog speech synthesiser in block diagram form

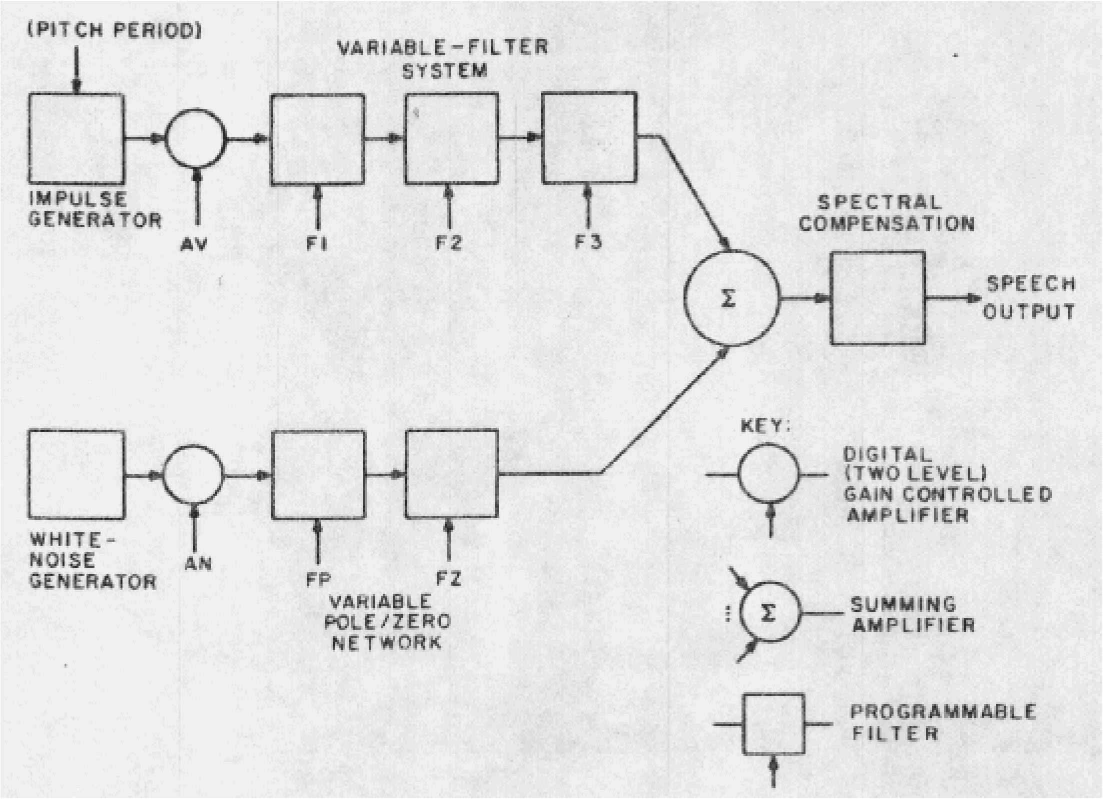
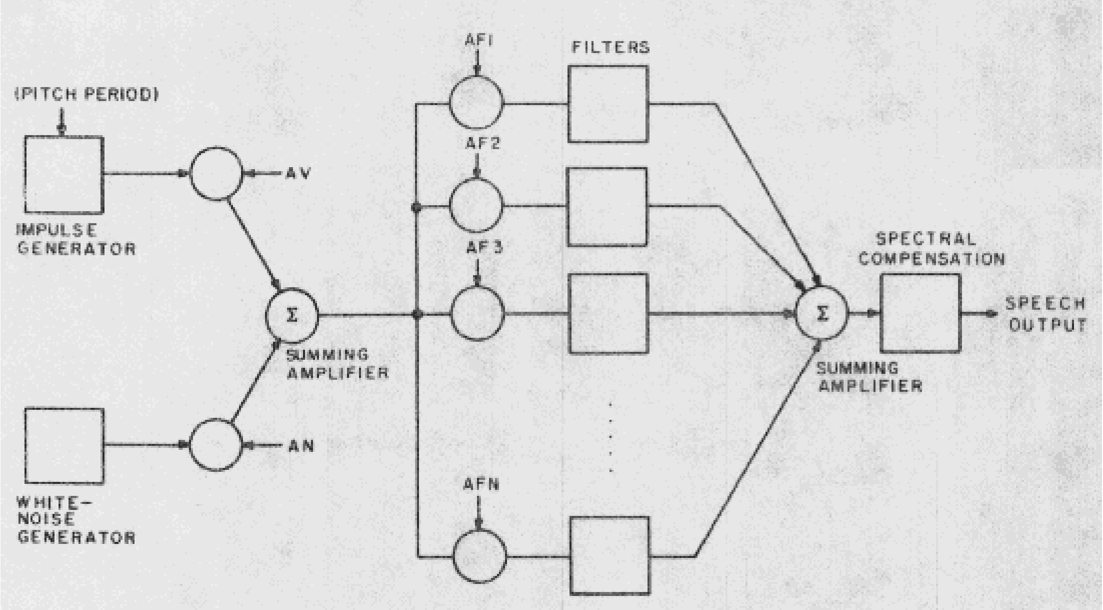


Figure 2:The parallel analog speech synthesiser in block diagram form



To produce speech, a separate circuit, or combination of circuits, must be provided to generate each of the above classes of phonemes.
Among possible reaIizations of such a synthesizer there are the serial analog and parallel analog forms. Figure 1 illustrates a block diagram of a serial analog design,

and figure 2 shows the general organization of a parallel analog synthesizer.

The parallel analog method was the realization chosen by Ai Cybernetic Systems for its synthesizer module. The parallel realization was chosen because of the low digital information transfer rate and the smaller number of bits recquired to control the filters which simulate the resonant cavity of the vocal tract.

In the Ai Cybernetic Systems design, the rush of aspirared air is generated by the NOlSe of a zener diode operated at its knee, amplified many times, as shown in figure 3. The action of the larynx is simulatedl by an integrated circuit function generator. One or both of these circuits is selected to produce the excitation necessary to generate any one class of phonemes. The actual phoneme perceived is determined by the duration of the excitation and the selected formant filters. Figure 4 shows the typical formant filter circuits which are digitally activated by analog switches.

The control of the several analog switches is provided by a read only memory which is addressed by the ASCII bit patterns ideritified in table 1.

No hard and last rules exist in the design of tIe circuitry to generate a phoneme. In fact small changes in component values can make large differences in the phoneme which is actually heard. Because no set rules exist, a steady stream of listeners must parade before the machine while it is being designed in order to determine which phoneme the synthesizer is really saying.

The phenomenon of "tired ears" rapidly sets in; and a person will begin, after a bit, hearing any one speech sound as a whole array of possible phonemes. Suggestion, on the other hand, is an ever obtuse enemy to the designer. Surprisingly, almost any speech sound can be suggested to sound like a great number of alternate phonemes, especially after 20 to 30 minutes of intense listening.

Once the design is experimentally determined, careful procedures must be followed to insure that when the circuit is duplicated, it produces each phoneme properly. This means precision components must be used, as small changes in values can make the difference between moderately distinct speech and a fairly mushy speech.

Analog simulation of the vocal tract is the only method of true speech synthesis known. A popular alternate method of speech production (actually, reproduction) is the storage of digitized speech in a ROM. When the stored information is clocked out of the ROM at the proper rate and smoothed by a low pass filter, the generated speech can be quite clear and distinct. But it is important to note that this is not synthesized speech. In effect, this method is no different than any other method of recording speech. Yet, the method does have the advantage of producing readily understood words by a computer or calculator. However, the vocabulary is totally predefined and must remain small due to the high cost of storing this kind of generated speech. Moreover the repertoire of this kind of speech is limited to the person who initially spoke the recorded words.

Figure 3: The excitation sources of the Ai Cybernetic Systems Model 1000 Speech Synthesizer.
The rush of air through the vocal passages is simulated in the upper branch while the action of the larynx is simulated in the lower branch.

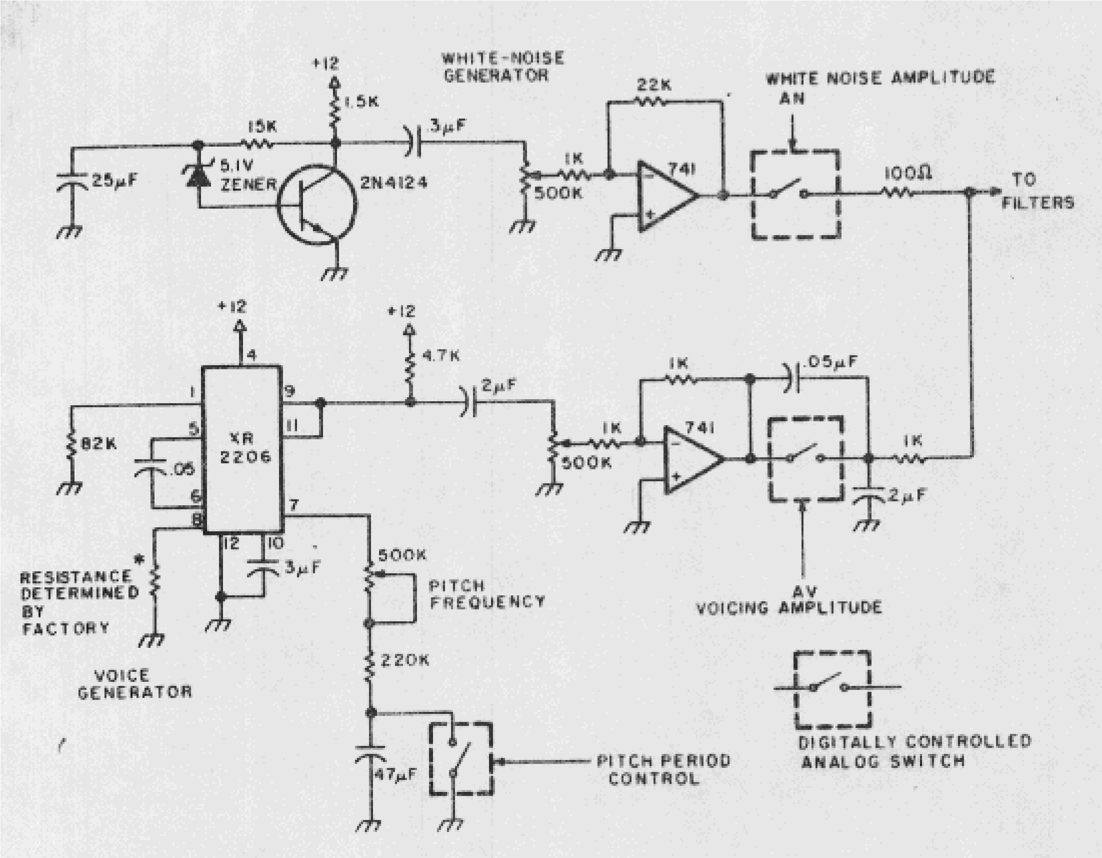


Figure 4: The parallel filter network of "Model 1000. The filter frequencies and quality factors chosen depend on the number of filters used to divide the voice frequency spectrum. Ideally, the center frequencies of the filters should lie some-where near the commonly occurring formant frequencies.

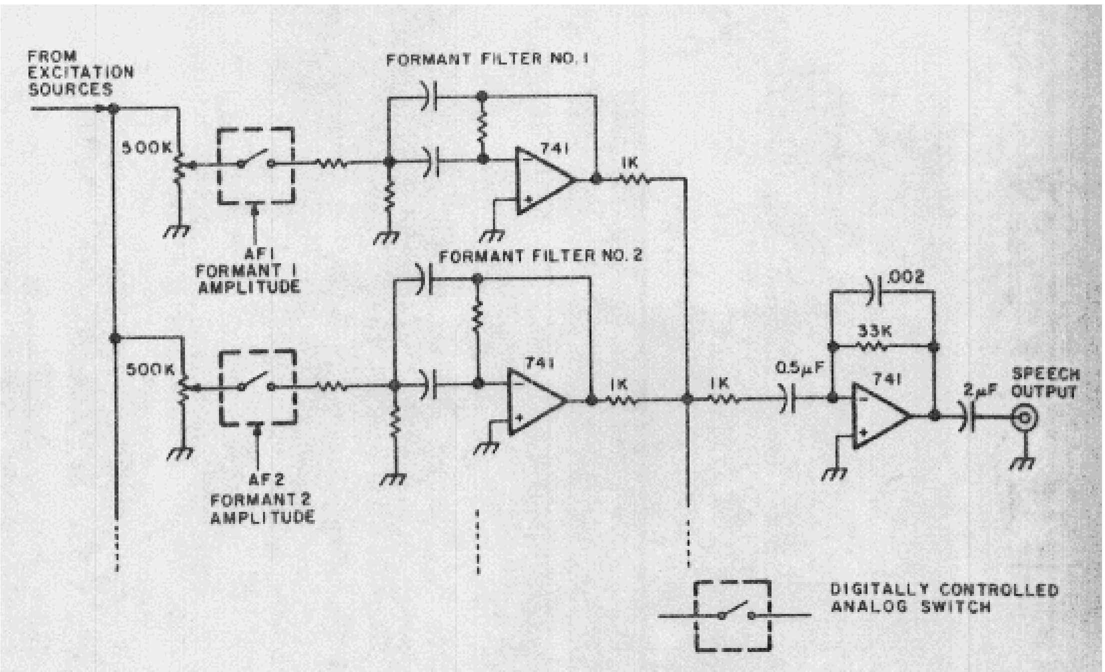
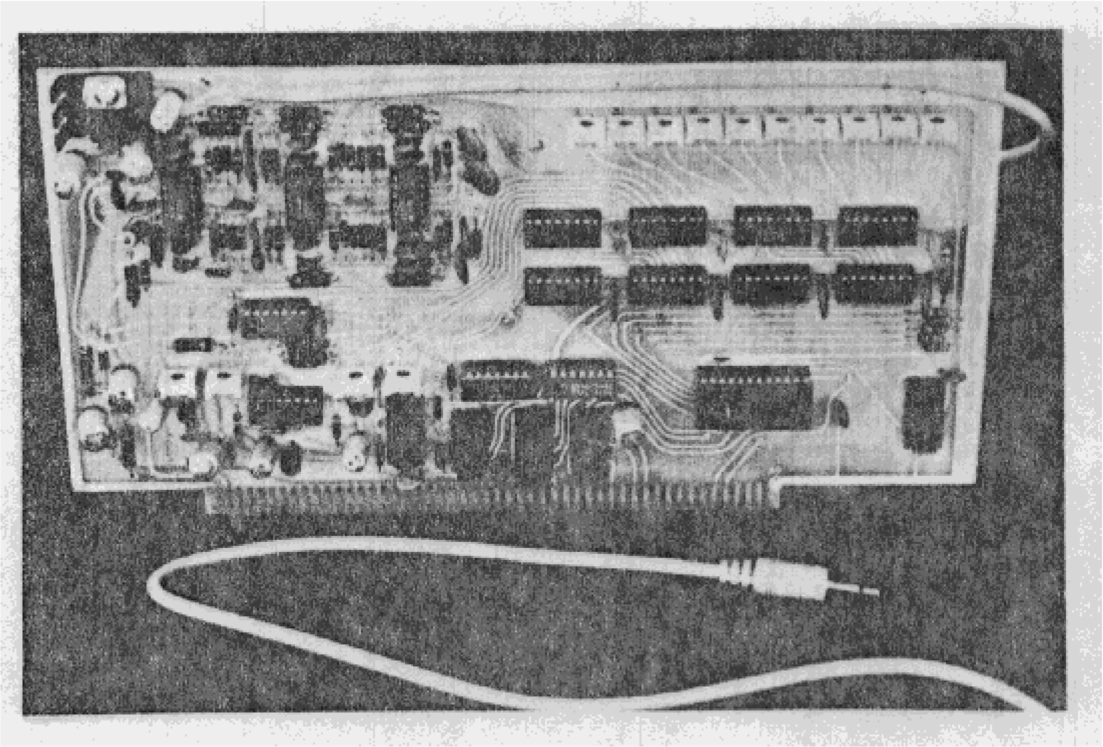


Photo 1: The Ai Cybernetic Systems Model 1000 Speech Synthesizer. The synthesizer is primarily an analog circuit controlled digitally. Ten active filters composed of 15 operational amplifiers are mounted in the upper left corner he board. Directly beneath these resonant-cavity simulating filters are the vocal excitation circuits The right half of the

board is composed of the ASCII character decoding circuits and phoneme memories Four 32 x 8 ROMs control the 16 analog switches to select the proper combinatIon of circuits to generate any one phoneme. A device-busy flag is returned for the duration of the phoneme.



Synthetic speech, on the other hand, is generally not as clear and distinct The proper transitions from phoneme to phoneme, the automatic emphasis given to leading or terminating consonants, and the intonation of a rhythm in speech which is associated with a word's importance or placement, are all facets of human speech which are difficult to properly recreate in machine produced speech. The determination of accurate rules to account for these factors has been the subject of active and intense research at centers here, and in Europe and Japan, including Bell Telephone Laboratories, the Haskins Laboratories of New York, the Royal Institute of Technology in Sweden, and the Musashino Electrical Communication Laboratory in Tokyo. On the whole, totally satisfactory rules have not yet been worked out although a great deal of progress has been made in the last 20 years. Machines which do incorporate the known rules quickly become elaborate and expensive (In the tens of thousands of dollars).

Simplified speech rules can be incorporated in a much smaller machine, but the burden of intelIigibility now falls upon the listener. The produced speech is not natural speech It sounds for all the world like the speech produced by the robots of 1950's grade science fiction movies. But it is intelligible and it is quickly learned. Because the machine pronounces every phonememin the same fashion each time it occurs, a listener quickly gains a feeling for the speech. The process is not unlIke learning to listen to a newly-arrived foreigner who possesses a strong accent.The fashion by which he mispronounces the English phonemes is quickly learned and intelligibility increases rapidly. The difference with synthetIc speech is that the speech is truly an alien form of speech, not often heard before by many of us.

As to the naturalness of synthetic speech, M D Mcllroy of Bell Telephone Labs wrote this in 1974 (in "Synthetic English Speech by Rule," Computer Science Technical report No. 14, Bell Telephone Laboratories:

The Computer Science Center at this laboratory has experimented with inexpensive speech synthesizer- (presumed to be the Votrax)as a regular output device in a general purpose computing system. Our intention was not to do speech research or to create artificial speech as an end in itself. In the present state of the art, those goals requIre much more elaborate facilities than we have at our disposal.

We wished to see what uses might evolve when speech became available more or less on a par with printed output. For this goal, "naturalness" was not a prerequisite, any more than it is for printed output. Most computers still print mainly in upper case, are incapable of printing mathematical notation, and normally produce cryptic codes or tabular stuff that require considerable indulgence to be understood. Since printed gobbledy-gook is so widely accepted from computers - and fed into them, witness any manufacturer's operating system manual - we suspected that spoken gobbledygook might be quite passable, too, except for one severe difficulty:

Being ephemeral, sounds must be understood at first hearing. As it turns out, long speeches are hard to understand, as are extremely short utterances of very simple words out of context. But given a little familiarity with the machine's accent, one finds short sentences to be quite inteligible.

The phonemes generated by the Model 1000 synthesiser appear in table 1. Each phoneme has been assigned an ASCII character to represent its particular sound. The assignment was done in the most intuitive manner possible; the consonants are generally the consonants as they appear on the keyboard, but there are many more vowels than a, e, i, o and u. Non-alphanumeric characters were chosen to represent the remaining vowels and consonants in such a manner that they could be easily associated with their sound. As examples of this the number symbol, "#" is used to signify the vowel er as in number, "&" for the vowel ae as in and "(" for ah and ")" for aw representing the position of the tongue when these vowels are spoken, "!" for the sharp stound of uh, "+" for the fricative consonant th as in thaw, and "/" for the sh in slash.

Table 1: List of Phonemes

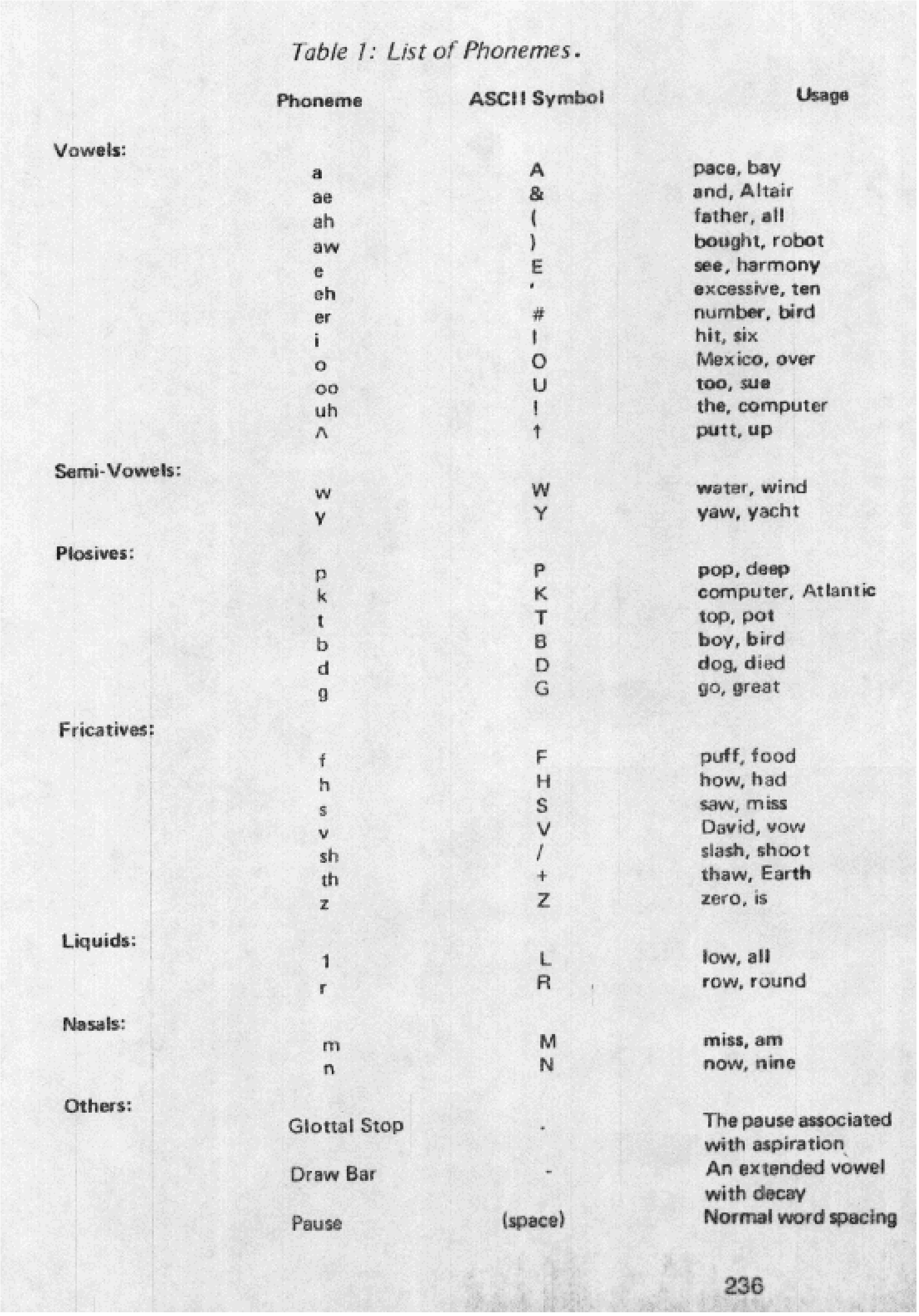


Table 1: List of Phonemes.

	Phoneme	ASCII Symbol	Usage
Vowels:			
	a	A	pace, bay
	ae	&	and, Altair
	ah	(	father, all
	aw	)	bought, robot
	e	E	see, harmony
	eh	'	excessive, ten
	er	#	number, bird
	i	I	hit, six
	o	O	Mexico, over
	oo	U	too, sue
	uh	!	the, computer
	^	↑	putt, up
Semi-Vowels:			
	w	W	water, wind
	y	Y	yaw, yacht
Plosives:			
	p	P	pop, deep
	k	K	computer, Atlantic
	t	T	top, pot
	b	B	boy, bird
	d	D	dog, died
	g	G	go, great
Fricatives:			
	f	F	puff, food
	h	H	how, had
	s	S	saw, miss
	v	V	David, vow
	sh	/	slash, shoot
	th	+	thaw, Earth
	z	Z	zero, is
Liquids:			
	1	L	low, all
	r	R	row, round
Nasals:			
	m	M	miss, am
	n	N	now, nine
Others:			
	Glottal Stop	.	The pause associated with aspiration
	Draw Bar	-	An extended vowel with decay
	Pause	(space)	Normal word spacing

The Model 1000 accepts a string of ASCII characters as if it were a normal printing device. Read only memories on the board convert the incoming ASCII symbol into specific control information which in turn determines the vocal source, duration and frequency content of the spoken phoneme. Less than 50 bytes of machine code or 8 lines of the typical BASIC are all that is required to generate a subroutine to accept a string of characters and output it character-by-character to the synthesizer. For example, to write the phrase "I am a talking robot", on a printer or display peripheral, an ASCII character string is set up and sent to the output device. In BASIC, if C$ is the argument of the output subroutine, the set up would be; C$="I AM A TALKING ROBOT."

To have the synthesizer say the same phrase, the setup for the phonetic output routine with argument P$ might be: P$= "&IE AM AE T).. KEN- RO.B). .T" (The ASCII symbols are taken from table 1.)

The long vowels I and A occur in this passage. As a rule, most of the long vowels are not really vowels at all but rather diphthongs composed of a sequence of pure vowels. Pronounce out loud each of the phonemes in the phrase above, referring to table 1 as necessary. Remember that each phoneme has only one specific sound. Playing the part of a synthesizer yourself, you will find that you can say any English word with the phnemes of table 1.

Programming the Model 1000 synthesiser is easy once you actually begin to listen to what you say and learn to rely less on how a word is written. English is a hodge podge of languages and carries with it all the alternate symbolisms of the pronunciations of its root languages. Purely phonetic languages such as the Polynesian languages of Samoa or Tonga could be made to spoken almost as they are written. This is unfortunately not true of English; homonyms such as "won" and "one" and "two", "too" and "to" abound.

Generally, only one phonetic spelling exists for any one word regardless of the number of alternate written spellings. It becomes important to identify the sounds that you actually are saying when a word is pronounced. The word "one" is phoneticized using the phonemes of table 1 as W!N in similarity to the word "won";"two" is programmed as TOU- more as if it were the written word"too". For most Americans, there is no difference in the way these words are pronounced.

Proceeding in the same fashion, the remaining numbers up to ten are typed in as:

T+#E- FO#- F&IE..V SI..KZ S'-VIN AE..T N&IEN T'N

Again, pronounce these phonetic spellings to yourself. As you will discover, phonetic spellings are quickly deduced and learned. In a very short period of time, it becomes possible to make the machine say anything.

At that point, conversational computing takes on a whole new meaning. Interactive computing will never again be the same once your computer has actually spoken to you.

Bibliography

1. Speech synthesis, Benchmark Papers In Acoustics, 1973, J.L. Flanagan and L.R. Rabiner, es. Dowden,Hutchinson and Ross, Stroudsburg PA. A Collection of the best papers on speech synthesis over the past 35 years.
2. Synthetic voices for Computers, 1970 J.L. Flanagan, C.H. Coker, L.R. Rabiner, R.W. Schater, N Umeda in IEEE Spectrum 7:22-45. An authoritative overview of the speech synthesis procedure
3. The Synthesis of Speech, 1972. J.L. Flanagan, Scientific American 226:48-58. A simplified work of the IEEE Spectrum article above
4. IEEE 1974 Speech Recognition, Proceedings 1974, L. Erman, ed. IEEE, NY. A bit too technical for a first introduction but a good measure of where things are going.

COMMERCIAL PRODUCTS

At the present time, two speech synthesizers are both commercially available and affordable by the hobbyist. One is the Votrax produced by:

Vocal Interface Division
Federal Screw Works
500 Stephenson Dr
Troy MI 48084
Price, approximately $2,000
Interfacing: Parallel or Serial (RS-232)

The second is the Model 1000 manufactured by:

Ai Cybernetic Systems
PO Box 4691
University Park NM 88003
Price, $425
Interfacing: Electrically and mechanically compatible with Altair/IMSAI/Poly-88 bus structure.

Either company will be pleased to provide literature free of charge. A demonstration tape is available from Ai Cybernetic Systems for $5 and a complete programming guide, theory of operation manual and phonetic glossary is available for $2.50.