The Chatterbox

A simple speech synthesizer for demonstration and amusement

by Ian H. Witten, M.A., M.Sc., Ph.D., M.I.E.E. **and Peter H. C. Madams,** B.Sc., M.Sc.

Department of Electrical Engineering Science, University of Essex.

The device described is a hand-controlled, electronic model of the acoustic properties of the vocal tract, and was built to illustrate the physiological and acoustic nature of speech. Although designed as a portable demonstration and lecturing tool it makes a fascinating toy for adults and children alike, and has been used as a stimulus for retarded and autistic children. After discussing the nature of speech and the mechanism of electronic speech synthesis, the authors explain the design principles of the Chatterbox and in a later article will give further circuit details and instructions on how to make it talk.

PEOPLE SPEAK by using their vocal chords as a sound source, and making rapid gestures of the articulatory organs (lips, tongue, mouth, etc.). The resulting changes in shape of the vocal tract allow the production of the different sounds that we know as the vowels and consonants of ordinary language. For several years it has been possible to simulate the action of the vocal tract electrically, using a device similar to an electronic organ to produce sounds with the same character as those of human speech. The first of these "speech synthesizers," built in the early 1950s, comprised many racks of equipment, consumed a lot of power, and cost a great deal of money. Now, however, with the advent of cheap integrated circuits, it has become possible to build simple, compact, and quite inexpensive synthesizers, without sacrificing the ability to produce the full range of speech sounds.

Of course, to make the ever-changing patterns of speech, a synthesizer needs some form of continuously varying control, and just as there are many vocal tract organs involved simultaneously in speaking, so it is necessary to control several parameters of the synthesizer at once. Most speech research laboratories nowadays use a digital computer to manipulate the control signals for their synthesizers. However, for the purposes of informal experiments with speech or just to learn about the sounds we make, a pair of hands will suffice – with the added advantage that the operator can use his long-standing experience with real speech to mould the sounds into voice-like ones.

By way of illustration of these points, we have built a small, manually controlled speech synthesizer, suitable for home construction – the "Chatterbox" (Fig. 1). In experienced hands it can be encouraged to utter recognizable words and phrases ("hello," "how are you," etc.), while even a complete novice can make it generate a great variety of astonishingly different noises, all of which are immediately recognizable as speech-like. The Chatterbox was originally designed as a portable demonstration and lecturing tool for illustrating the different sounds of speech and how they can be synthesized, but we quickly found that the fascination of artificial speech makes it a successful and compelling toy for adults and children alike. As a hand-controlled, electronic model of the acoustic properties of the vocal tract, it provides a natural feel for the growing science of phonetics — the study of what people do when they are talking and when they are listening to speech.

This first article discusses the nature of speech and the mechanisms of electronic synthesis. The Chatterbox design is described later, with some circuit details, as a concrete example of the implementation of a speech synthesis system.

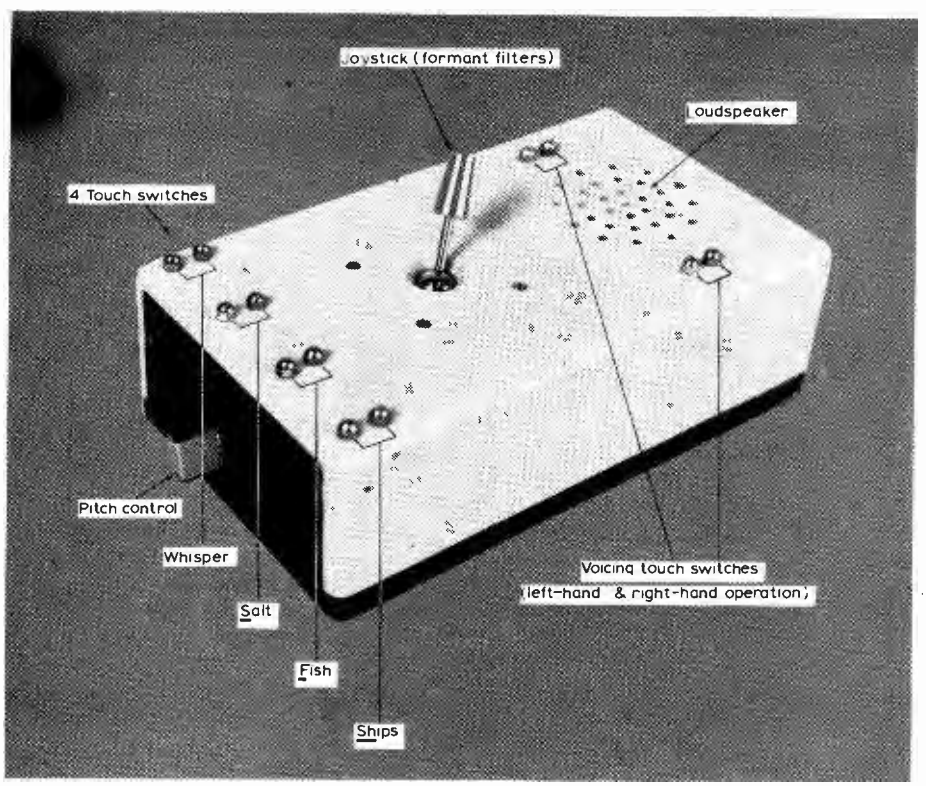


Fig. 1. *The Chatterbox, showing the controls for hand operation.*

The anatomy of speech

The so-called "voiced" sounds of speech – like the sound you make when you say "aaah" – are produced by passing air up from the lungs through the larynx or voicebox, which is situated just behind the Adam's apple. The vocal tract from the larynx to the lips acts as a resonant cavity, amplifying certain frequencies and attenuating others.

The waveform generated by the larynx, however, is not simply sinusoidal. (If it were, the effect of the vocal tract resonances would merely be to give a sine wave of the same frequency but amplified or attenuated according to how close it was to the nearest resonance.) The larynx contains two folds of skin – the vocal cords – which blow apart and flap together again in each cycle of the pitch period. The pitch of a male voice in speech varies from as

low as 20Hz to perhaps.250Hz, with a typical median value of 100Hz. For a female voice, of course, the range is correspondingly higher. The flapping action of the vocal cords gives a waveform which can be approximated by the triangular pulse of Fig. 2. This has a rich spectrum of harmonics, decaying at around 12dB/octave, and each harmonic is affected by the vocal tract resonances.

A simple model of the vocal tract is an organ-pipe-like cylindrical tube with a sound sourcé at one end (the larynx) and open at the other (the lips), as shown in Fig. 3. This has resonances at wavelenghts $4L$, $4L/3$, $4L/5$, . . . , where L is the length of the tube; and these correspond to frequencies $c/4L$, $3c/4L$, $5c/4L$, . . . Hz, where c is the speed of sound in air. Calculating these frequencies, using a typical figure for the distance between larynx and lips of 17cm, and $c = 340$ m/sec for the speed of sound, leads to resonances at approximately 500Hz, 1500Hz, 2500Hz, . . .

When excited by the harmonic-rich waveform of the larynx, the vocal tract resonances produce peaks known as *formants* in the energy spectrum of the speech wave (Fig. 4). The lowest formant, called formant 1, varies from around 200Hz to 1000Hz during speech, the exact range depending on the size of the vocal tract. Formant 2 varies from around 500 to 2500Hz, and formant 3 from around 1500 to 3500Hz.

Of course, speech is not a static phenomenon. The organ-pipe model describes the speech spectrum during a continuously held vowel with the mouth in a neutral postion such as for "aaah." But in real speech the tongue and lips are in continuous motion, altering the shape of the vocal tract and hence the positions of the resonances. It is as if the organ-pipe were being squeezed and expanded in different places all the time. Say "ee" as in "heed" and notice how close your tongue is to the roof of your mouth, causing a constriction near the front of the vocal cavity.

Linguists and speech engineers use a frequency analyser called a sound spectrograph to make a three-parameter plot of the variation of the speech energy spectrum with time. Fig. 5 shows a spectrogram of the utterance "go away." Frequency is given on the vertical axis, and bands are shown at the beginning to indicate the scale. Time is plotted horizontally, and energy is given by the darkness of any particular area. The lower few formants can be seen as dark bands extending horizontally, and they are in continuous motion. Notice that in the neutral first vowel of "away," the formant frequencies approximate the 500Hz, 1500Hz and 2500Hz that we calculated earlier. (In fact, formant 2 is around 1250Hz and formant 3 around 2300Hz.) The fine vertical striations in the spectrogram correspond to single openings of the vocal chords. Of course, the pitch is continuously changing throughout an utterance, and this can be seen on the spectrogram by the differences in spacing of the striations. Pitch change, or *intonation*, is singularly important in lending naturalness to speech.

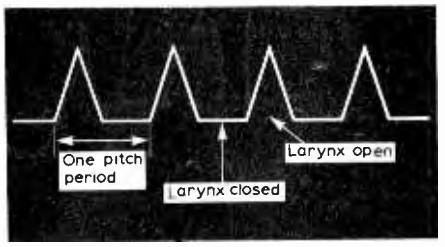


Fig. 2. *Approximate waveform produced by the larynx.*

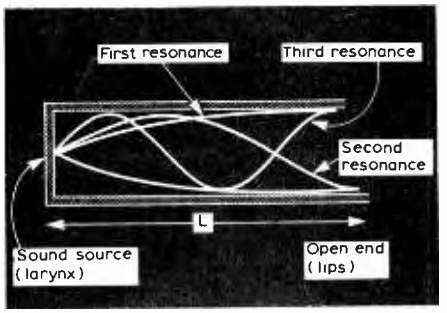


Fig. 3. *Resonances in the organ-pipe model of the vocal tract.*

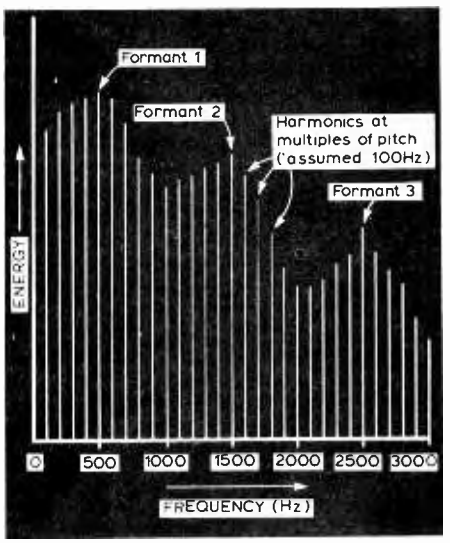


Fig. 4. *The energy spectrum of speech, showing three formants.*

Fig 5. *Spectrogram of the utterance "go away."*

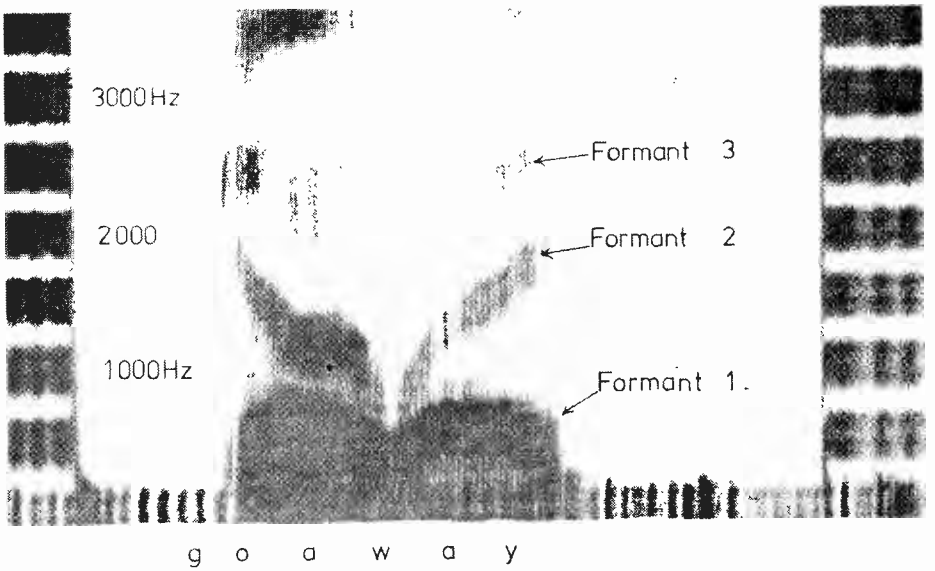


Table 1. The vowels and their formant frequencies

Vowel name	*Example of use*	*F1 (Hz)*	*F2 (Hz)*
UH	ab(ove)	500	1500
A	bud	700	1250
E	bed	550	1950
I	bid	350	2100
O	bod	600	900
U	good	400	950
AA	bad	750	1750
EE	bead	300	2250
ER	bird	600	1400
UU	brood	300	950
AR	bard	700	1100
AW	board	450	750

On a spectrogran, a continuously held vowel shows up as a static energy spectrum. But beware – what we call a vowel in everyday language is not the same thing as a "vowel" in phonetic terms. Say "I" and feel how the tongue moves continuously while you're speaking. Technically, this is a *diphthong* or slide between two vowel positions, and not a single vowel. And there are many more phonetically different vowel sounds than the a, e, i, o and u that we normally think of. The words "hood" and "mood" have different vowels, for example, as do "head" and "mead." The principal acoustic difference between the various vowel sounds is in the frequencies of the first two formants. Table 1 gives a list of the English vowels, with a one- or two-character name for each, an example

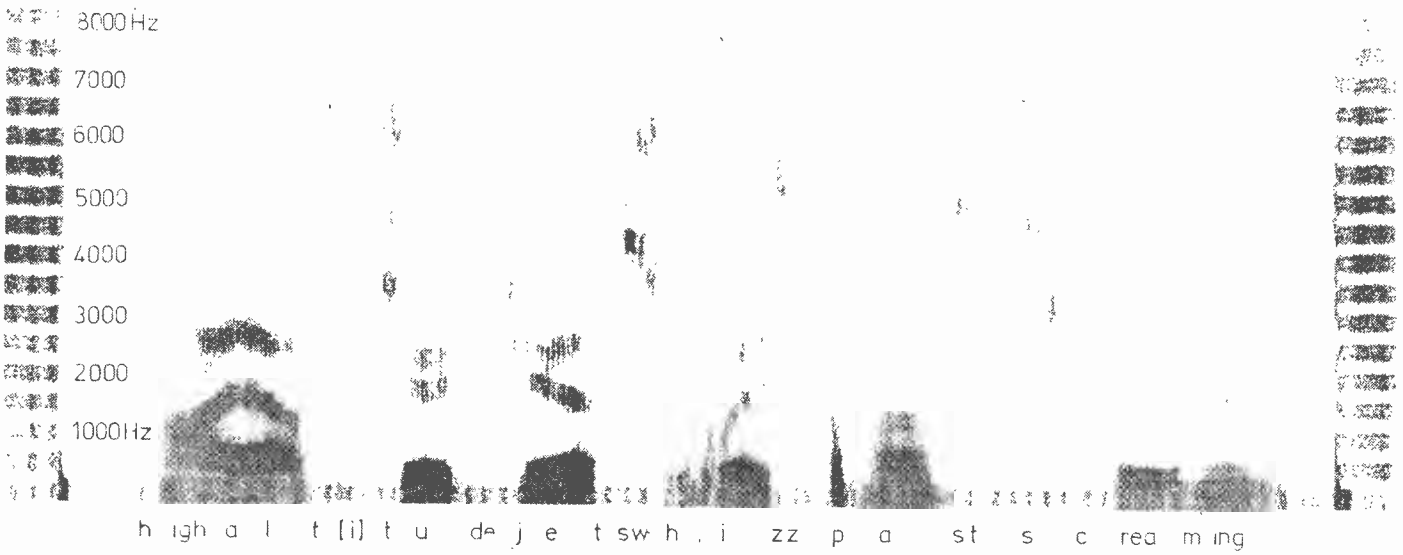


Fig. 6. *Spectrogram of "high altitude jets whizz past screaming."*

word, and the two formant frequencies which characterize the sound.

Speech involves other sounds, different from the voiced ones that we have been discussing so far. When you whisper, the folds of the larynx are held slightly apart so that the air passing between them becomes turbulent, causing a noisy excitation of the resonant cavity. The formant peaks are still present, superimposed on the noise. Such "aspirated" sounds occur in the "h" of "hello," and for a very short time after the lips are opened at the beginning of "pit."

Constrictions made in the mouth produce hissy noises such as "ss," "sh," and "f." For example, in "ss" the tip of the tongue is high up, very close to the roof of the mouth. Turbulent air passing through this constriction causes a random noise excitation. For "sh," the tongue is flattened close to the roof of the mouth, in a position rather similar to that for "ee" but with a slightly narrower constriction, while "f" is produced with the upper teeth and lower lip. If the larynx is vibrating as well we get the corresponding voiced sounds "z," the "zh" in "azure," and "v." Because they are made near the front of the mouth, the resonances of the vocal tract have little effect on these hissy sounds. The complicated acoustic effects of noisy excitations in speech can be seen in the spectrogram Fig. 6 of "high altitude jets whizz past screaming."

Speech synthesis and synthesizers

The idea of artificial speech has always fascinated man. The first genuine talking machine appears to have been demonstrated in 1791 by one Baron von Kempelen, who used bellows to inject sound into a leather tube which modelled the vocal tract, and was deformable with the hands to imitate the different vowel sounds. Progress continued sporadically (one notable achievement being Alexander Graham Bell's encouraging his pet dog to talk by manipulating its vocal tract by hand while the dog growled), until the need for bandwidth reduction for efficient use of communication channels in the 1940s and 1950s stimulated serious research on the acoustic nature of the speech signal. Since then, the advent of widely-available real-time computing power has encouraged work on speech synthesis under computer control, and the difficult problems of pronunciation, speech rhythms, and intonation are currently being tackled to exploit this novel and effective computer output medium.

In order to simulate electrically the resonating action of the vocal tract on the sound generated by the larynx, a waveform generator and several resonant filters in cascade are needed. Varying the frequency and amplitude of the sound source simulates changes in the pitch and loudness of the speech, and different vowels can be made by adjusting the positions of the resonances appropriately.

Further analysis of the organ-pipe model reveals that simple second-order resonators with unity d.c. gain are appropriate filters, with the slightly unusual requirement that the bandwidths should remain constant as the resonant frequencies are altered, producing sharper resonances at higher frequencies (Fig. 7). The phase response of the filters is not important. Such resonators can be achieved with simple active filter circuits.

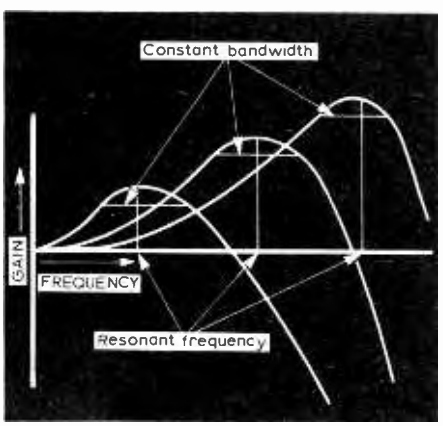


Fig. 7. *Amplitude profiles of the resonant filter.*

Although vocal tracts, like organ pipes, have an indefinite number of resonances, in practice only a few filters are employed in the chain (Fig. 8). Most existing synthesizers simulate four or five formants, of which typically only the first three have controllable resonance positions. In fact, two formant filters are sufficient to generate most vowel-like speech sounds: a third is especially useful in distinguishing the "r" in "rice" from the "l" in "lice." Omitting the higher resonances means that some compensation filter needs to be introduced to give spectral lift at higher frequencies.

Whispery sounds can be synthesized by injecting noise into the chain of formant filters, instead of the harmonic-rich pulse of the waveform generator. For the sibilant sounds made at the front of the mouth, the noise should not be injected into the formant chain, but instead passed through a separate high-pass resonance whose centre frequency can be controlled to give the sounds "f," "sh," and "ss."

These considerations lead to the block diagram of Fig. 9. A synthesis system similar to this was invented in 1951 by Walter Lawrence, and he called it PAT – Parametric Artificial Talker. The eight circled numbers represent parameters of the system, and if they are varied appropriately, it can be persuaded to give a respectable imitation of almost any speech utterance. For example, the parameter tracks for "six" are shown in Fig. 10 as a set of eight graphs. You can see the onset of the

hissy sound at the beginning and end (parameter 5), and the amplitude of voicing (parameter 1) come on for the "i" and go off again before the "x." The pitch (parameter 0) is falling slowly throughout the utterance.

Naturally, storage of the parameter tracks presents some problems. In the earliest version of PAT, eight parameter-versus-time graphs were painted on a glass slide, which was scanned photo-electrically to read off the parameter values. Lawrence was fond of disconnecting the pitch parameter and controlling it directly with a potentiometer. One of the utterances for which he had prepared a glass slide was "What did you say before that?" and he could manipulate the pitch by hand to change the emphasis, getting "What did you *say* before that?", "What did you say *before* that?" and so on. Even artificial singing proved possible! In fact, the earliest computer conversation on record was between PAT and a Swedish synthesizer, called "Ove." The inventors, Gunnar Fant and Walter Lawrence, stood by their machines on a stage at an international acoustics meeting, Fant with a small transistorized table-top box, and Lawrence with several great racks of valve-based equipment. The conversation went

Ove: "How *are* you?"
PAT: "What did you say before *that*?"

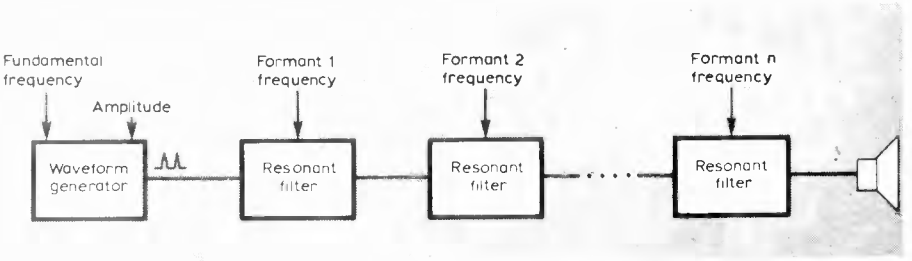


Fig. 8. *Simulating the resonance action of the vocal tract.*

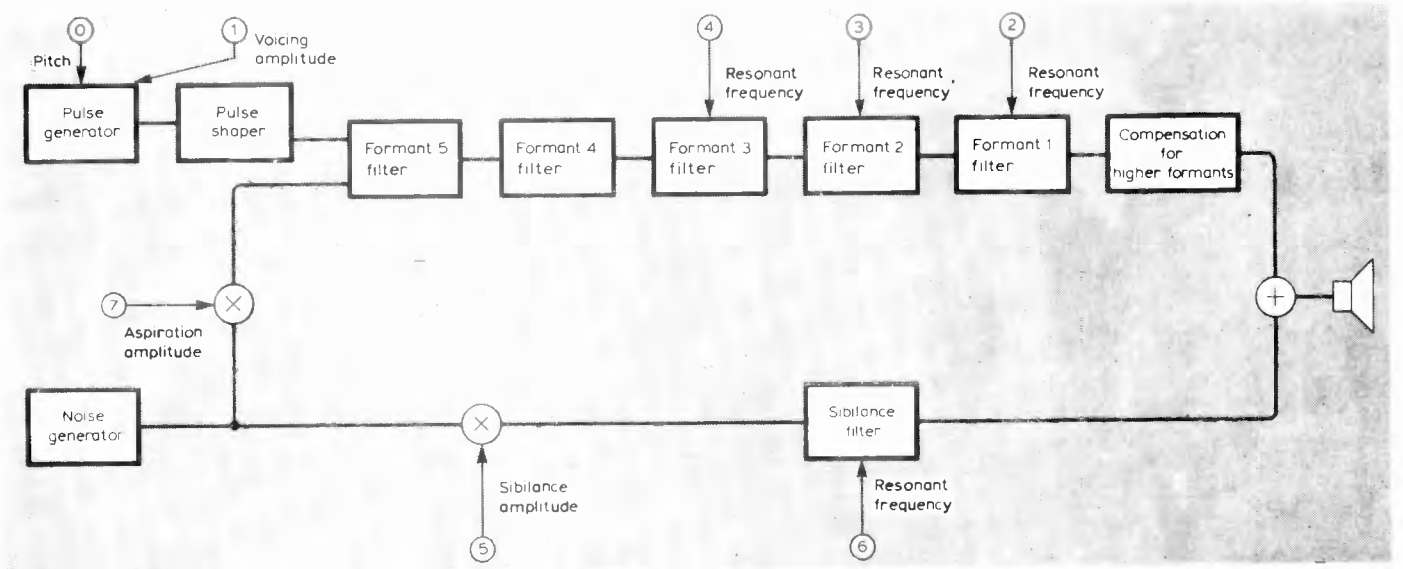
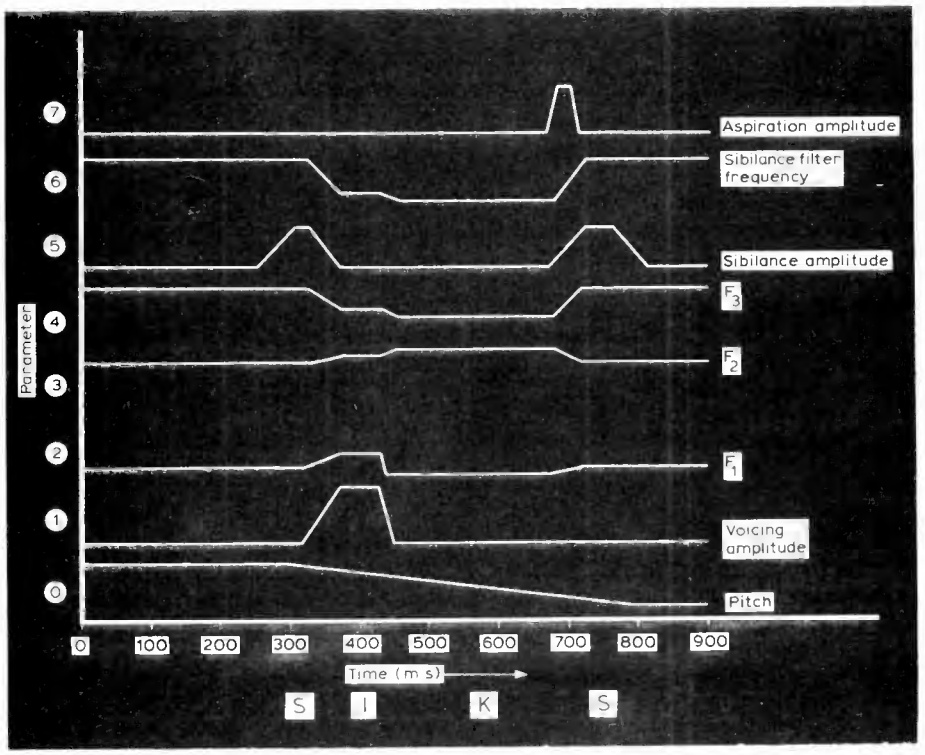


Fig. 9. *Block diagram of PAT, the Parametric Artificial Talker.*

Fig. 10. *Parameter tracks for "six" with the Parametric Artificial Talker.*



Ove: "I love *you*."
PAT: "What did you say *before* that?"
Ove: "I love *you*."
At which point, PAT burst into song (no prizes for guessing the words!).

To obtain good speech, the best way of getting parameter tracks is to derive them from spectrograms of human utterances. Although this is a tiresome and time-consuming process, it gives the synthesizer a chance to reproduce the precise acoustic quality of the original speech. However, the parameter tracks of Fig. 10 are stylized: they don't come from a human utterance. In fact they were generated by a computer programme from the input "S I K S", a phonetic transcription of the word. This programme has direct control over the parameters of a hardware synthesizer through a computer interface, and will attempt to speak any utterance that is entered in phonetic format. In practice, the most difficult parameter to control in a convincing way is pitch. The intonation of speech is subtle, and evades classification into a form that a computer programme can handle.

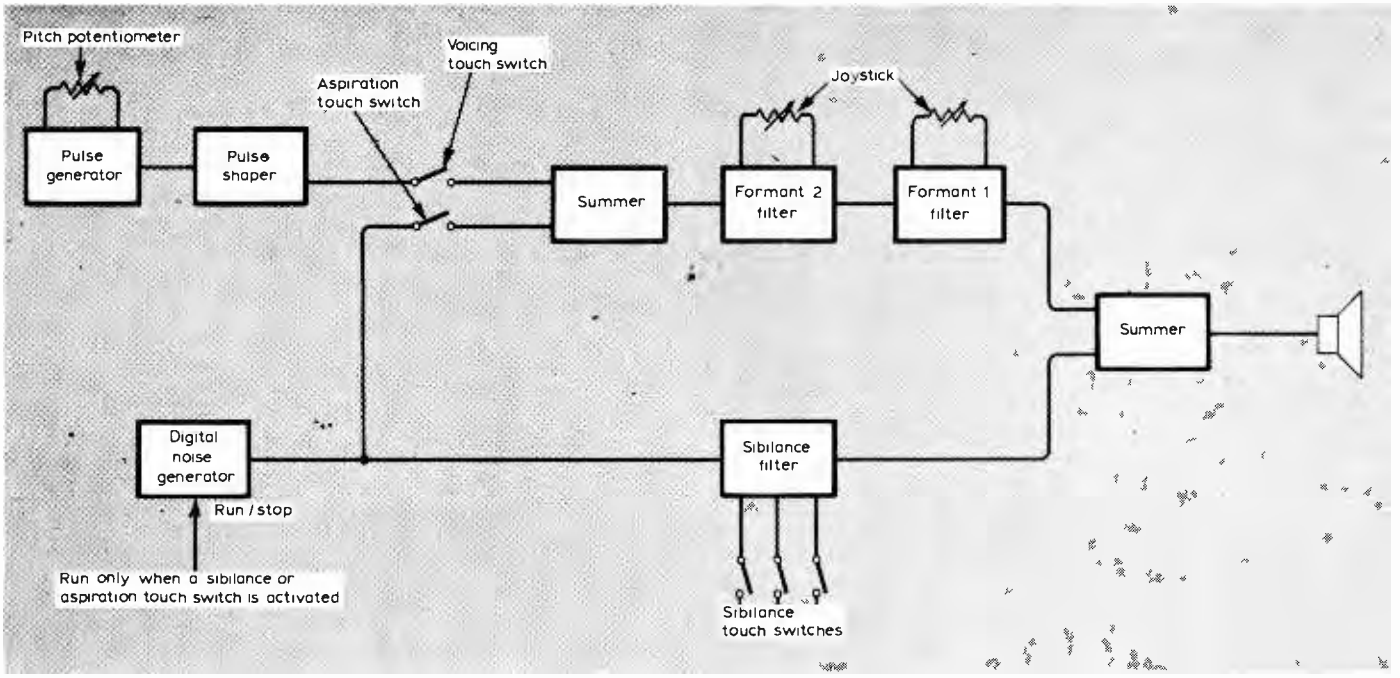


Fig. 11. *Chatterbox block diagram.*

The Chatterbox, however, avoids these difficulties of computer control by using a person to manipulate the parameters. Generating natural-sounding intonation is easy for people, and this turns out to be true even if they have to use their hands rather than their vocal tracts to control the pitch.

System design of Chatterbox

The manual controls. Hands were never intended to speak! In your vocal tract, separate muscles control a multitude of parameters of the system simultaneously in order to produce speech sounds. Pitch is controlled by the vocal cord tension, amplitude by the lung pressure, and vowel quality by the many dimensions of movement of the tongue, teeth and lips. The greatest challenge of the Chatterbox design was the engineering of the man-machine interface: it is difficult physically to find enough degrees of freedom to control it with the hands. In fact, we even considered using arm and leg movements in addition to hands, but felt that these detracted from the neatness and compactness of the toy.

A single X-Y control is used to vary the two formant filter frequencies. Two models of Chatterbox have been designed and constructed, one with a joystick control and the other with a stylus and resistive plastic pad instead. Fortunately, the recent popularity of quadraphonic audio systems means that it is quite easy to get hold of a compact joystick assembly designed as a quad balance control. Ours had two tracking potentiometers for each direction of motion, and we took advantage of this in designing the formant filters. The alternative stylus arrangement gives a two-dimensional position indication by injecting a current from the stylus tip into a uniformly resistive plastic sheet, and monitoring the current from the sides of the sheet. Since it is fairly difficult to lay hands on a suitable resistive sheet for the stylus model, and the circuitry to take advantage of this is more complex anyway, we will describe only the joystick version here.

The pitch is varied by a potentiometer. We decided, after some experimentation, that linear rather than rotary control feels more natural, so a slider potentiometer is used. This is operated with one hand while the other directs the joystick.

Turning to control of the volume of the sound, it transpires that contrary to intuitive expectations, it is not important to provide variation of the amplitude of the voice source, apart from the obvious necessity to switch it on and off. Different vowel sounds do have different amplitudes, of course, because of different degrees of mouth opening (compare the vowels in "mad" and "mood", for example). However, this is taken care of by the formant filters: resonances for "mood" will naturally produce a weaker sound than in "mad" because they occur at lower frequencies, and the constant bandwidth property of the filters gives them less amplification (that is, lower Q) at lower resonance frequencies. The amplitude of the sound produced by the vocal cords corresponds more to vocal effort than to loudness, and this is more or less constant for the great majority of speech sounds. Hence we use a simple switch to turn the voicing on and off.

The hissy sounds pose the most difficult control problem. Aspiration (whispering) can be treated just like the voicing amplitude: we need only be able to turn it on and off. The joystick formant control can then be used to whisper different vowel sounds. However, constrictions in the front of the mouth must be treated separately, for here the sound type ("ss", "sh" and "f") needs to be controlled independently of the voicing, so that the counterparts "z", "zh" and "v" can be produced simply by superimposing a normal vowel-like sound. We opted for separate switches for these three noises, instead of an analogue control which would simulate the tongue positions more accurately. These are operated by the same hand that manipulates the pitch potentiometer, as is the aspiration switch. This arrangement is not parti-

Further reading

The anatomy of speech

The "source-filter model of speech production," which separates the sound source (larynx) from the filtering operations of the vocal tract, was treated most comprehensively by Gunnar Fant in *Acoustic theory of speech production*, 1960.

The sound spectrograph was developed in 1946 by Koenig, Dunn, and Lacey ("The sound spectrograph," *Journal of the Acoustical Society of America*, vol. 18, pp. 19-49) and is described, with hundreds of spectrograms, by Potter, Kopp and Green *(Visible Speech*, 1947).

Speech synthesis and synthesizers

A classic book is *Speech analysis, synthesis and perception*, by James Flanagan of Bell Laboratories in the USA (1965, revised 1972). There is a book of collected papers on speech synthesis by Flanagan and Larry Rabiner called *Speech synthesis* (1973). A British contribution is *Speech synthesis* by John Holmes of the Government Joint Speech Research Unit (1972). Walter Lawrence wrote "The synthesis of speech from signals which have a low information rate" (in *Communication theory*, edited by W. Jackson, pp. 460-469) when he invented PAT in 1953 in the Government Signals Research and Development Establishment.

cularly easy to use, but since pitch control is unimportant during hissy sounds, it is possible to share the pitch hand satisfactorily between all these functions.

It is essential, however, that the voicing on/off switch is easy to operate while complicated pitch movements are being made, because the moment of onset and offset must be timed precisely, without disrupting the smooth flow of intonation. In the pad-and-stylus Chatterbox, it is possible to detect electrically when the stylus is in contact with the pad, and this is used to turn on the voice. A similar arrangement could be made in the joystick model if the act of grasping the joystick were detected, but this would involve modification of the joystick assembly. We opted instead to site a switch contact where it could easily be reached with the heel of the hand that operates the joystick.

All switches on the Chatterbox are touch switches, and work by detecting the skin resistance when two adjacent contacts are touched together. These are much easier to use – and cheaper to build! – than, say, pressure operated microswitches.

The overall system. The Chatterbox is essentially a simplified version of Lawrence's original PAT. As shown in Fig. 11, it consists of two parallel signal paths, a voicing/aspiration path (top of diagram) and a sibilance path (bottom). Two formant filters form the upper path, each controlled by one direction of the joystick. These can be excited by a simulated larynx pulse, produced by a variable-frequency impulse generator, or by a noise source (for aspiration). We use a digital pseudo-random generator implemented by a feedback shift register with exclusive-OR feedback. The same noise source drives the lower sibilance path, which includes a high-pass resonance to give the noise an appropriate colouration. The position of the resonance is controllable to three places by touch-switches. A full circuit diagram of the system is shown in Fig. 12, and the various sections of this will be explained here and next month.

Source for voiced sounds, and mixer. A simple circuit using c.m.o.s. gates forms the basis of the voicing waveform generator. This circuit is used because of its simplicity and low cost. The 100kΩ potentiometer varies the frequency of the oscillation. The output from the oscillator is inverted, delayed, and "ANDed" with itself. This produces a train of spikes which is the required harmonic-rich voicing waveform. The circuit is turned on and off by a logic signal from one of the touch switches shown in Fig. 12. The period can be adjusted from a very low value, 20Hz (good for sound effects and "creaky voice"), up to 200Hz. The harmonic content of the spiky waveform is very high and provides a suitable excitation for the formant filters. The 741 amplifier that follows this circuit is used both to add a signal from the noise generator to produce aspirated sounds and to adjust the amplitude of both sources.

To be continued