VOCAL SYNTHESIS

by Peter Hillen
consultant, National Semiconductor

An article on computer generated speech wouldn't seem right without a reference to Star Trek or 1984. That was it. The fact is that computer speech is here today and is in reach of electronic music performers both from an availability and a cost stand point.

The first speech synthesizers were a mechanical replica of the vocal tract. Soon after followed the electronic synthesizers using analog circuitry. The problem with analog synthesis is that sound sequences are hard to generate. Next came the digital synthesizers. Entirely digital speech synthesis techniques require a large, high speed and expensive computer to perform all the algorithms necessary to produce speech. The breakthrough comes in the form of a trade off in digital vs. analog technique which reduces computer speed, power and cost with the addition of some special purpose analog hardware. One such speech synthesis sytem which uses this approach is the Computalker (PO Box 1951, Santa Monica, Calif. 90406). The Computalker speech synthesizer is a card full of analog electronics which fits into computers designed to be compatible with the S-100 hobbyist computer interface for which there are a number of suppliers of inexpensive ($500) good quality computers.

The Computalker breaks down the task of speech synthesis into two parts. First is the generation and formating of analog signals to make speech-like sounds. The Computalker circuits take care of this. Second is the generation of control signals to sequence the Computalker module in the correct way to make its output sound like speech. This is accomplished by the computer using a digital interface to the module and extensive software to follow the complex rules of speech.

Let us examine how we humans speak and break the contributing functions into blocks which the computer can control. For reference, a schematic of the whole vocal synthesis system is shown in Figure 1. It is the electronic analog of the vocal tract which starts at the vocal cords and ends at the lips. As we go through the derivation you will find that the electronic counterparts of our vocal tract are made up of the very familiar electronic music functions of voltage control.

First, the vocal cords or vocal folds. The vocal folds open and shut one end of the vocal tract interrupting the flow of air from the lungs. These short bursts of air serve as the oscillator in our vocal system. The form of these waves looks like a half wave rectified sine wave as is shown in Figure 2. The electronic analog of the vocal folds is a variable frequency oscillator. It is variable because the frequency at which each of us use our vocal folds is different. Furthermore, when we speak, the frequency of the sound generated by the vocal folds does not stay constant. Its variation is used to increase the meaning of the words being spoken. Speech amplitude is also important to meaning, adding emphasis to what is being spoken. The amplitude of the oscillator is controlled through a variable gain amplifier.

The sound generated by the vocal folds makes its way up the vocal tract to the lips. The vocal tract can be treated as a straight tube. From physics, it is known that a characteristic of a tube such as this is to be resonant at odd multiples of the wave length of the tube. From the analysis of speech, it has been found that only the first three of these resonances need be considered when developing a model for speech. If you have looked into your mouth lately, you may have noticed it doesn't look like a straight tube. All along the way, there are things which can change the characteristics of the sound. Some to consider are the shape of the mouth, the nasal cavity, the position of the tongue, the teeth and the lips. Each of these can be modified depending on what is being spoken to deliver the desired sound characteristics. The resonant frequencies are not fixed at the odd multiple intervals and can vary in accordance with the change in the physical shape of the tract. These resonances are simulated in Figure 1 as three variable frequency filters in series with the vocal cord oscillator and amplitude control. This path is called the oral

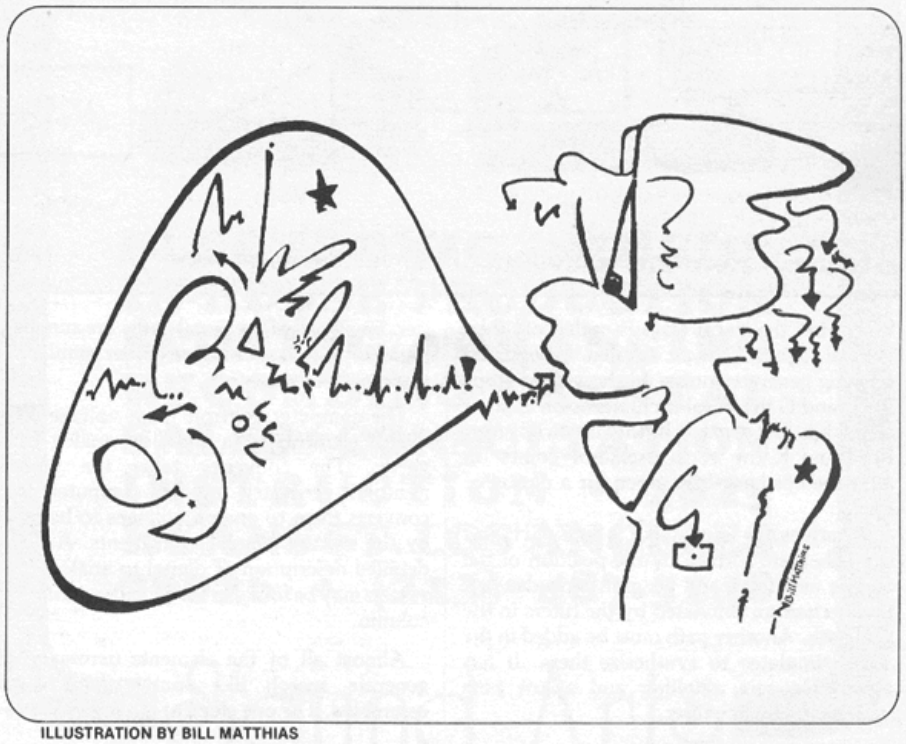

ILLUSTRATION BY BILL MATTHIAS

VOCAL SYNTHESIS

path. It is used to form all of the vowel sounds (A, E, I, O, and U) and some of the consonant sounds.

Consonants are divided into 4 classes: liquids, stops, fricatives and nasals. The liquids are comprised of the sounds W, Y, R and L. They are formed very much like vowels except that the way the mechanism is used is timed differently.

The next two classes, stops and fricatives, make use of noise. We generate noise by exhaling through a narrow opening formed by the lips and tongue. Noise is simulated in the electronic vocal tract by a noise source followed by variable gain amplifiers to control its amplitude.

Stops come in two forms: voiced and unvoiced. They are both formed in a similar way. First the vocal tract is sealed off then opened slightly. The air exhaled through this opening generates noise. In the voiced stops B, D, and G the vocal oscillator is on as soon as the lips are opened. In the unvoiced stops P, T and K the vocal oscillator comes on after the lips have been open for a period of time.

Fricatives are sounds like S, F, Z, SH, and TH. They are formed by the position of the tongue and teeth and are not dependent on the mechanism simulated by the filters in the oral path. Another path must be added in the vocal simulator to synthesize them. It has a variable gain amplifier and a low pass variable frequency filter.

The last path in the vocal simulator must be added to take into consideration the nasal cavity. The nasal cavity is important in generating consonant sounds like M and N. They are made by closing the vocal tract and exhaling through the nose. The cavity itself is fixed in size and can be simulated by a fixed bandpass filter of fairly wide band width.

Finally every block in Figure 1 has been defined. All along filters, amplifiers, oscillators and noise sources have been mentioned. Let's be specific and relate these to electronic music modules. The vocal cord oscillator is nothing more than a VCO with about a 2¾ octave range and a modified sine wave output. The noise source is a standard wideband white noise type. The amplitudes of both of these signal sources are controlled by VCAs which contour the amplitude envelope before passing the signal on to the filters. There are two types of filters used in the vocal simulator: variable and fixed. Both are the bandpass type. The format filters ae variable and work like a VCF. The nasal cavity filter is fixed and works like a preset graphic equalizer. Finally all of the signal paths are summed together by a summing amplifier similar to a microphone mixer.

The computer controls these voltage controlled elements by a digital to analog converter. The converter takes the binary numbers generated by the computer and converts them to analog voltages to be used by the voltage controlled elements. A more detailed description of digital to analog converters may be found in the Synapse computer column.

Almost all of the elements necessary to generate speech like sounds have been assembled. The one element missing is the cohesive force to link everything together in an intelligent way to generate speech. This force comes in the form of a computer software program. The software program is responsible for the timing of the changes in amplitude and frequency of all the elements involved. The problem now is how does a human operator of the computer give the commands to synthesize speech? There are two methods.

The first gives excellent speech quality but is time consuming to generate. It is referred to as synthesis by hand. Synthesis by hand is accomplished by using a spectrum analyzer and amplitude envelope follower to determine the frequency/amplitude profile of the speech for a human speaking into a microphone connected to the equipment. The data from these devices is usually in the form of a graph which must be analyzed to define each parameter on an instant by instant basis and then coded into the computer. Additional editing must be done to improve the quality of the speech by compensating for deficiencies in the hardware. Synthesis by this method requires expensive analysis equipment and is restricted by the fact that the synthetic speech is patterned after the voice of the person making the sample.

The second method which is most useful is the use of a phonetic alphabet. This is a set of symbols compatable with the keys on a typewriter and corresponds to the elements of sound (vowels and consonants) described above. The phonetic symbols contain information about timing, intonation and stress levels for the words being synthesized. A person familiar with the dictionary can translate directly from English text to phonetic text. The computer takes the phonetic text, interprets the rules and produces speech.

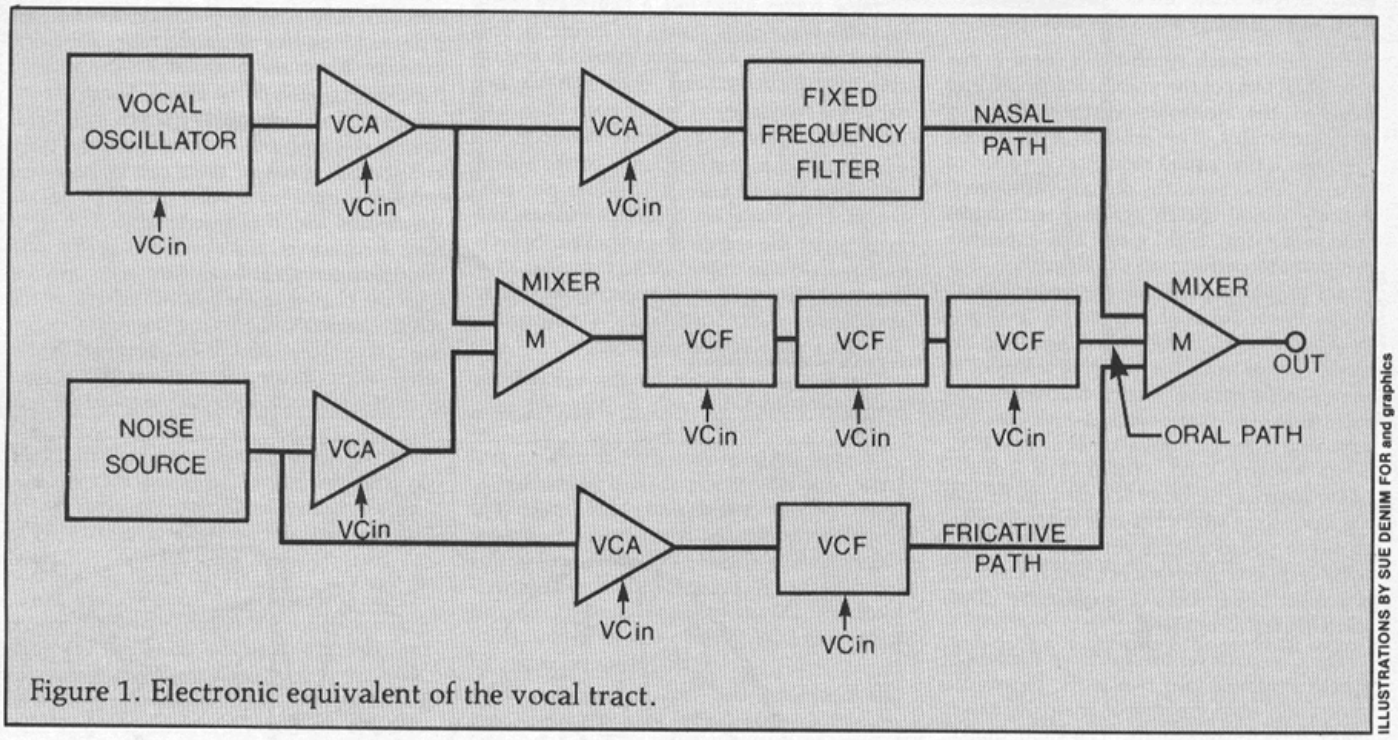

Figure 1. Electronic equivalent of the vocal tract.

ILLUSTRATIONS BY SUE DENIM FOR and graphics

Figure 2. Output of vocal oscillator.

Take as an example the word *this*. The phonetic dictionary representation is *dhih3s/* which seems pretty obscure.

It's not really all that bad. The *DH* symbolizes the *the* sound as in the word *that*. *Th* symbolizes the *i* as in the word *bit*. The *3* is a stress level symbol which follows a vowel. *S* is the symbol for the sound an *s* makes in the word *sat*. With a little memorization and practice it is not difficult to compose sentences quickly. The power of the phonetic dictionary is not without penalty. English is a very complicated language and is full of many exceptions to the rules. Words like *two*, *too* and *to* which are spelled differently but sound alike (all are synthesized as *tuw*) prevent a one to one relation between English text and phonetic text. Further problems arise with context dependent words like: "*read* Synapse" (*riyd*) or "yesterday I *read* Synapse" (*rehd*). Even more important is the relationship of each phonetic element in a sentence to the others. There must be a smooth transition from one to the next. At this time the stage of development of the computer software for all speech synthesizers is not good enough to provide it. Because of these reasons phonetic dictionary generated speech is not as easy to understand as hand generated speech. Good example of both hand generated speech and phonetic text speech are on a demonstration cassette from Computalker available at the address given above for $2.95.

A brief overview of synthetic speech has been given. Now the question is how can the electronic musician use it. Robbie the Robot like sentences tucked away in the middle of a composition spouting some profound idea (ALL PERSONS ARE SYBLINGS) would be a waste for such a tool. A microphone into a fuzzbox or use of the EMS voice synthesizer could accomplish a similar effect much easier. The real value of a computer speech synthesizer comes from what it can do that the human voice can't.

In the most generalized case the speech synthesizer can be used as a computer controlled sequencer having nothing to do at all with speech. The VCO for the vocal cords can be tuned over about a 2¾ octave range and the filters, amplifiers and noise generator can be used in a similar manner to ones in a conventional synthesizer. This is useful in storing repeating patterns such as bass lines or rhythmic sounds. It allows for easy recall because the information can be preprogrammed and stored for later use.

Now for the human sounds. Two parameters the speech synthesizer can control better than we humans is frequency and time. First consider time. Computer speech differs from tape recorder modified speech with respect to the effects of speed changes. The tape recorder can double or half the speed at which something is said but at the same time alters the frequency of the speech. The speech synthesizer does not. Voice frequency stays constant no matter how fast or slow a word is said. It is even possible to vary the speed within a word or sentence on a phrase by phrase basis. For example, one could hold the *a* in the word *pause* indefinitely, certainly longer than a human voice could. Also words or groups of words could be synthesized backwards. This is extremely easy to do when compared to the old method of tape editing, permitting real time backward speech on a word by word basis. An interesting case would be to have each word in a sentence followed by its mirror image in time.

So much for becoming unstuck in time. Frequency can also be manipulated in several ways. As voice frequency is independent of time in speech synthesis so is time independent of frequency. Sentences can be spoken at normal speed but the voice frequency could be an octave higher, a fifth lower or any place you want it. This opens up possibilities in a multi-track recording studio environment to perform near perfect multipart harmonies with exact timing and phrasing.

Voice frequency can be made independent of phonetic phrasing but would require a departure from phonetic dictionary. It is possible to have the vocal cord oscillator do the melody while the rest of the vocal tract circuits are "mouthing" something different. Taking this one step further, why not eliminate the vocal cord oscillator altogether and use alternate sound sources such as a guitar or recorded voices which would be modulated by the vocal tract circuitry? Such an effect would be similar to the now infamous blow bag used by Peter Framton. It would require modification to the speech synthesizer because of the vocal cord wave shaping filters.

As you can see the speech synthesizer opens up a vast new area of exploration for the electronic music composer and musician. The technology is new but does exist and is affordable to anyone who wishes to try. ^^^