Speech Synthesis

National Semiconductor
Application Note 252
Jim Smith
Dave Weinrich

INTRODUCTION

Electronic speech circuits offer a new dimension of sophistication to many modern machines. As annunciators in trains, elevators, office buildings, autos, airplanes, terminals, toys and games, etc., electronic speech circuits provide a more direct and natural announcement than bells, buzzers or lights. With electronic voice signals, complex directions can be clearly given in any language and with a minimum of effort.

In the past, electronic announcement systems required elaborate tape mechanisms. These systems were expensive and troublesome, so their use was limited to the small number of applications that required speech announcements (e.g., telephone announcement systems). The first all-electronic systems used analog to digital conversion techniques to convert actual voice into digital signals. These digital speech signals were then stored as PCM or delta modulation signals in semiconductor memories. The major problem with this arrangement was the massive memory required for a moderate amount of announcement time. One second of digital speech, in this configuration, required from 16k to 100k bits of memory.

The latest solution to electronic speech is known as speech synthesis. This technique provides a dramatic reduction in the memory required for one second of speech. Memory requirements range from 400 bits to 2000 bits per second depending on the desired speech attributes and overall quality. The synthesizer techniques takes advantage of the fact that speech signals are highly redundant and predictable. By coding only the slowly varing coefficients of speech or by dramatic compression of digitized speech, significant bandwidth reductions in the digitized signal can be realized. These techniques, coupled with LSI semiconductor technology, make true voice synthesis practical.

The National Semiconductor speech processor chip (SPC) provides the complete speech synthesis reproduction circuitry needed to generate high quality and natural speech (male, female or a child's voice), eletronic tones or music. A complete chip set is called the Digitalker®. It consists of the speech processor chip and a speech ROM. The applications for this chip set are endless, but to name a few:

Voice interactive computer terminals

Automotive, nautical and aeronautical instrumentation annunciators

Voice-back units for banking, weather and time announcements, answering machines, etc.

Elevators, trains, subway systems, etc.

Consumer appliances, toys and games

Warning systems for fire and police emergency

All of these applications benefit from the lower overall costs, high reliability, excellent performance and fast control response afforded by the National Semiconductor Digitalker system. The remainder of this note will be devoted to a description of the MM54104 SPC, the technique used to synthesize speech and finally a review of the major Digitalker applications.

SPEECH SYNTHESIS

The basic phonological element of speech is the phoneme. The phoneme represents a simple sound that by itself cannot distinguish different words. Phonemes, together with speaker inflection, volume, emphasis, etc. are the fundamental building blocks of speech. The overall quality of any speech synthesizer, therefore, is directly controlled by its ability to faithfully reproduce all of the necessary speech attributes and not just phoneme reproduction.

The common American English language consists of approximately 38 to 40 phonemes—14–16 vowel sounds and 24 consonant sounds. Each phoneme is generated with either a voiced sound, as in "eye", or an unvoiced sound like the sh in "shy". This difference between a voiced and unvoiced sound is very important because the unvoiced sounds are generally fewer in number and less dependent upon the physiological characteristics of the speaker. A speech synthesizer, it turns out, can exploit this important difference. Finally, normal speech rates are approximately 10 to 15 phonemes per second (including silence intervals). Since 38–40 phonemes can be coded using 6 bits, the normal bit rate for phoneme reproduction is approximately 60 to 90 bits per second. This bit rate, however, contains only phoneme information which is only one of the many important speech attributes.

Since phoneme reproduction is a basic element in any speech synthesizer, an understanding of phoneme construction would be useful. Speech synthesis models use two driving functions, an impulse source for voiced sounds and a noise source (hiss noise) for unvoiced sounds. Each of these driving signals are filtered into specific frequency bands or formants by time-varying filters. The net result, for any particular set of valid filter coefficients, is a formant sound. In the human vocal tract system, the driving function consists of the lungs as the energy source, and the vocal cords for generating a voiced sound. The driving function for an unvoiced sound relies on the noise generated as air rushes through the vocal chambers and not on vocal cord vibrations. The formants are then generated by the resonant chambers of the throat, mouth and nasal cavities. By controlling the physical nature of these chambers with mouth position, tonque position and throat orifice size, a speaker can control the formants to generate a phoneme. It should be noted, however, that formants are identified by distinctive frequency bands. The unvoiced sounds do not generate these distinctive bands and therefore do not necessarily require the "normal" formants for a faithful reproduction. These sounds are characterized by a noise or hiss with very little resonance. This unvoiced resonance is normally identified as a fricative formant (e.g., the "sh" sound) and is characterized by an unusually large content of high frequencies.

ed with the nasal resonator. For the synthesis of a normal English vowel using a male voice, the three basic formants would fall into the approximate frequency bands of 200 Hz to 800 Hz, 900 Hz to 2300 Hz and 2400 Hz to 3000 Hz. The fricative formant is typically a pulse of high frequency noise in the band from approximately 2500 Hz to 8000 Hz, with the higher frequency fricatives like "th" usually much lower in relative amplitude when compared to the "sh" fricative sound.

The basic formant synthesizer requires formant filter coefficient data, amplitude control data and driving function control data. This minimum system could synthesize speech, but would not control inflection or emphasis. Its quality, therefore, can be very disappointing. Normal memory requirements for a minimal system are approximately 400 bits for one second of speech.

A second approach to speech synthesis does not automatically break speech into its minimum phonological elements. Instead, the speech waveform is sampled, digitized and compressed by the elimination of symmetrical redundancy and silent intervals, the use of adaptive delta modulation, and the adjustment of phase information in the digitized speech. In this way, speech elements can be synthesized as phonemes, phoneme groups, words or even whole phrases. Also, the attributes of the original speaker can be maintained if the synthesized elements are not broken down incorrectly (i.e., inflection can modify the sound of a phoneme if it occurs at the end of a word or phrase rather than at the beginning).

This approach is successful because unvoiced sounds have very few speaker defined characteristics. As a result, a relatively small set of unvoiced sounds can be used repeatedly.

This speech compression technique offers excellent quality at a low data rate. The synthesis of a male voice, using English, will usually require an average of 1000 bits per word. Because the technique can be applied to any voice frequency signal, it is also capable of synthesizing women's and children's voices, music and tones. This flexibility, plus the realistic quality of the synthesized speech, make this technique very attractive.

THE NSC SPEECH PROCESSOR CHIP (SPC)

The National Semiconductor speech synthesis system consists of the SPC device plus the speech memory (ROM) required to assemble a complete Digitalker kit. To this kit a customer must add a clock input signal or the necessary oscillator components, an audio filter and amplifier and the control circuit function. This would represent the minimum configuration shown in *Figure 1*. The maximum amount of directly addressable speech memory accessible by the SPC is 128k bits, but external page addressing by the control circuit function can increase this ROM field as required.

The SPC utilizes the speech compression synthesis technique. As mentioned earlier, this technique reduces the amount of memory needed to store electronic speech by removing the excess or redundant data from the speech

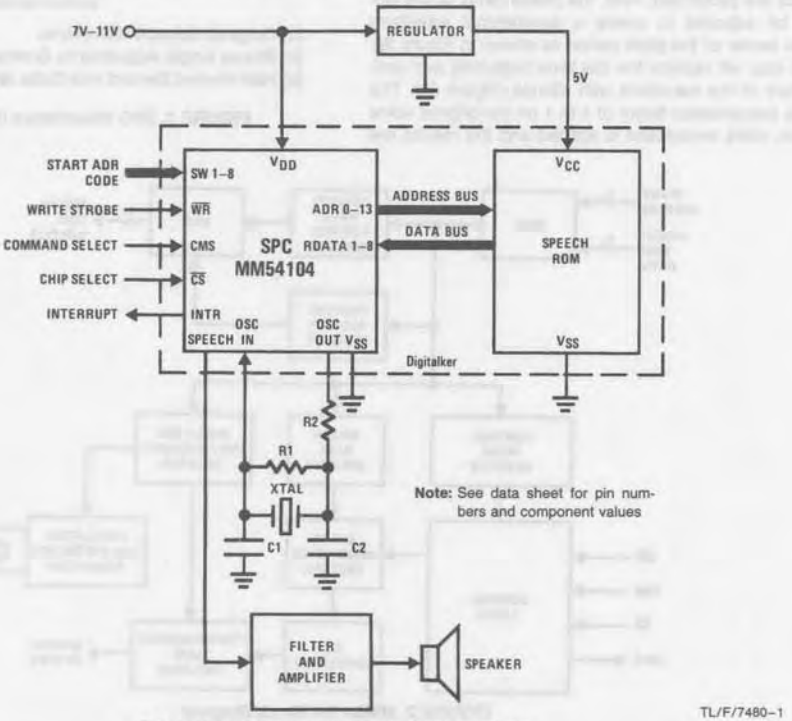


FIGURE 1. Digitalker Minimum Configuration

signal. The four main techniques to perform that task are:

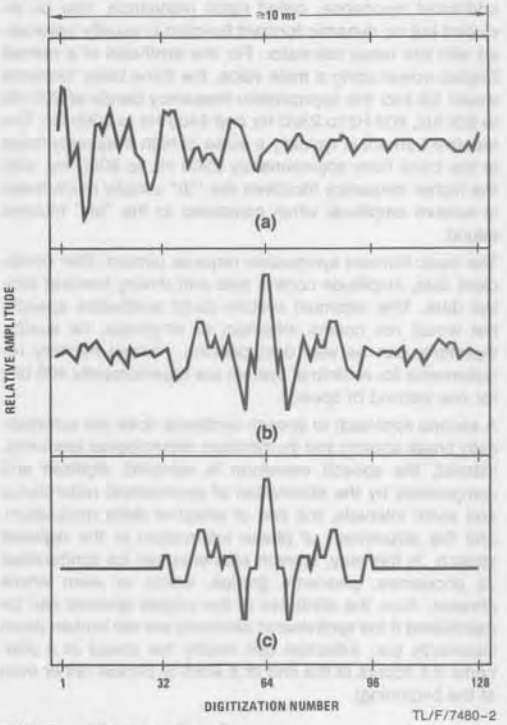
1. Elimination of redundant pitch periods
2. Adaptive delta modulation coding to minimize bandwidth and memory requirements
3. Phase angle adjustments to create mirror image symmetry
4. Replacing the low level portion of a pitch period with silence (half-period zeroing)

National Semiconductor uses an elaborate computer program to analyze a high fidelity tape recording and generate a ROM pattern that will faithfully synthesize the original voice message.

Figure 2 contains a block diagram of the MM54104 SPC. The eight-bit start address bus allows up to 256 separately defined sounds or expressions to be stored in the speech ROM. The control interface to the start address port can take the form of decoding logic, a MICROBUS™ port or mechanical switches.

When the $\overline{WR}$ goes high, the start address code is loaded into the control word address register. The SPC uses this control address to fetch the control word from ROM for the first block of speech data. The control word contains waveform information, repeat information and the address of the speech data. This address is loaded into the phoneme address register and is used to fetch the speech data used to recreate the speech waveform. Before the synthesis takes place, the waveform data must be decoded to provide information such as male or female, voiced or unvoiced, half-period zeroed or not half-period zeroed and silence.

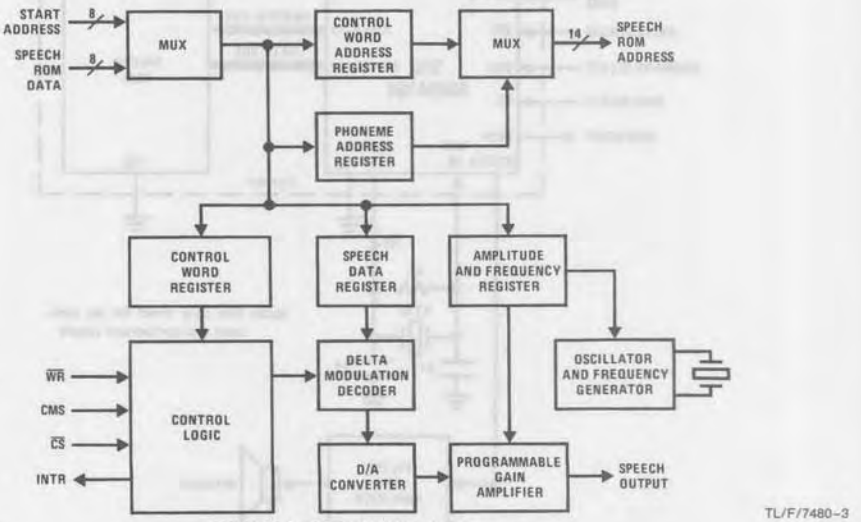
The unsynthesized waveform for a typical voiced pitch period might look like the signal shown in *Figure 3a*. In the process of converting this signal to a synthetic form, several operations are performed. First, the phase delay of the signal can be adjusted to create a symmetrical waveform about the center of the pitch period as shown in *Figure 3b*. The next step will replace the low level beginning and ending quarters of the waveform with silence (*Figure 3c*). The result is a compression factor of 4 to 1 on the original voice data. Now, delta modulation is applied and the results are shown in *Figure 3c*. Synthesis of the waveform starts with a period of silence (no speech data required), a quarter period of adaptive delta modulation-generated speech followed by



TL/F/7480–2

(a) Original Speech Waveform
(b) Phase Angle Adjusted to Create Mirror Symmetry
(c) Half-Period Zeroed and Delta Modulated

FIGURE 3. SPC Waveforms (After Mozer [2])



TL/F/7480–3

FIGURE 2. MM54104 Block Diagram

the same speech data fetched in reverse. Finally, the SPC will finish the last quarter cycle of the speech block period with silence. This phase modified speech data sounds the same as the original speech.

At the end of a waveform or speech block, the SPC makes a decision about repeating the sequence. Each waveform of a typical voiced signal may be repeated an average of 3 to 4 times. The typical unvoiced waveform may be repeated approximately 7 to 8 times. Once the proper number of repeats has been generated, the SPC will begin a new speech block sequence. This operation continues until the SPC has executed all control words associated with the original eight-bit start address code.

SPC speech signals are stored as adaptive delta modulation data. This encoding technique exploits the relatively predictible and slowly changing characteristic of voiced speech. Because of the small differential between successive speech samples, a delta value rather than an absolute value can be used to determine the actual speech signal. Addition of the delta value to previously accumulated values will result in a new output waveform signal level. An adaptive technique is used so that the delta step size can change in response to slope variations. This technique uses multiple delta modulation step sizes to obtain a more accurate resolution and yet, the required amount of stored data remains lower than the information required for a more conventional encoding scheme.

The internal SPC clock is derived from a programmable frequency generator. Variations in the frequency of this clock, through the control word, allow the SPC to add a rising and falling pitch to speech sounds and syllables. This derived pitch variation adds a natural inflection to the synthetic speech.

Just as pitch variations are used to increase realism, so must the SPC use gain variations. Both techniques are controlled by data stored at the beginning of a speech block and the programmable oscillator and output amplifier circuit blocks of the SPC.

Use of the Digitalker is quite straightforward and will be outlined in the next section. However, a point on application that must be covered in this note concerns the frequency response of the output speech. The ultimate quality of the Digitalker will strongly depend upon the filter, amplifier and speaker choices made by the user. For that reason, it is important to understand the output characteristics of the device.

Because the synthesized speech data is derived from a differentiated and sampled input signal, it is necessary to pass the output waveform of the MM54104 through a low-pass filter with a cutoff frequency of approximately 200 Hz and an attenuation characteristic of 20 dB/decade. This compensates for the high frequency pre-emphasis used in the synthesis technique. If the system of interest has a natural rolloff near 200 Hz, this low-pass filter can be eliminated. The important item is that the entire audio system should have a cutoff frequency of approximately 200 Hz. The placement of the cutoff frequency may be adjusted for the particular type of voice being synthesized. A low pitched man's voice might sound better with a 100 Hz cutoff point while women's and children's voices may show improvements with a 300 Hz cutoff. *Figure 4a* shows a filter and amplifier circuit for this minimum frequency response characteristic.

As an example of how the overall frequency response of a particular application can minimize the need for extra filtering, consider the Digitalker as a voice announcement circuit in a telephone system.

In this case, the telephone network provides a natural attenuation to high frequencies that balances the SPC high frequency pre-emphasis. As a result, the low-pass filter previously mentioned can be eliminated. However, because signal frequencies above 3 kHz must be attenuated before they are allowed to pass into the telephone network, a cutoff filter of 3400 Hz may be required in place of the previously mentioned 200 Hz low-pass filter. A good filter for this application is the National Semiconductor AF133 active filter.

In addition to the 200 Hz to 3400 Hz low-pass filter, an extra stage of filtering can be used for frequencies above 7 kHz. This filter is optional and is normally only used to further reduce sampling noise. Most systems can omit this filter, especially if the overall system bandwidth is not very wide. A second optional filter can be included to limit the overall low frequency response of the system. This high-pass filter would normally cutoff below 200 Hz (adjusted to match the 200 Hz low-pass if provided). This high-pass filter limits low frequency noise, and can usually be omitted if system characteristics do not require this function. A circuit having the full frequency response characteristic is shown in *Figure 4b*. *Figure 5* shows the recommended overall speech synthesis system frequency response.

APPLICATIONS

While the variety of synthetic speech applications are numerous, the actual implementation in any single application is usually limited to one of the following three techniques.

(a) Single channel, hardware control logic

(b) Single channel, software control logic

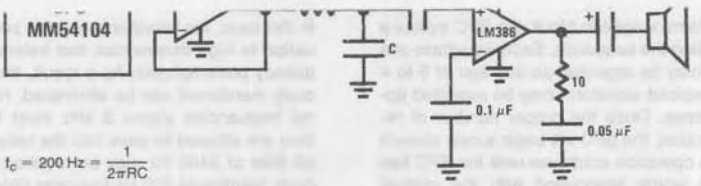
(c) Multichannel, hardware or software control logic

Each of these circuit approaches for the SPC will be discussed in this section. Particular emphasis will be placed on items (b) and (c), however, because of the broad application possibilities for these two techniques.

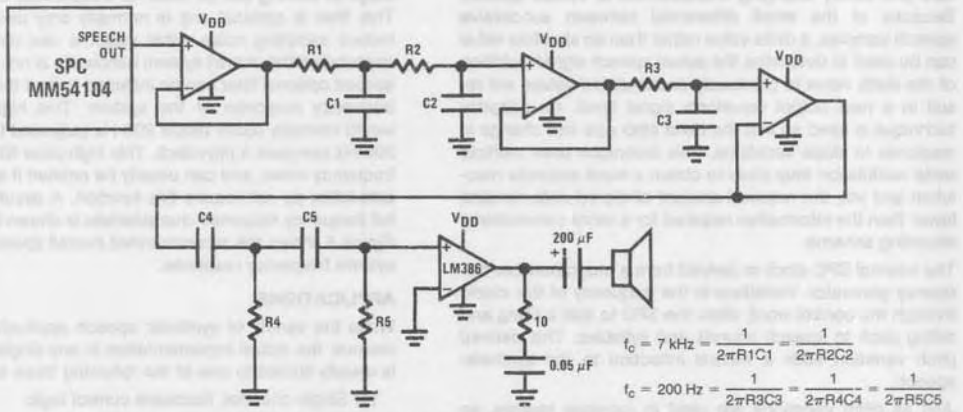
Certain applications require a relatively small number of sentences or announcements with very little similarity between the different sentences. An example of this application might be a talking elevator controller where the messages are brief and non-redundant (e.g., "going up, first floor, second floor", etc.). In this application, certain words are used repeatedly but the number of messages is limited and the length of each message is short. This application and others just like it, do not require the assembly of short phrases into complete sentences, nor do they require a dynamic message structure as would be required with an automatic bank teller (e.g., "your change is ten dollars") where a monetary amount may change from message to message. This fixed message application, therefore, may only require the minimum control circuit as shown in *Figure 6*.

In *Figure 6,* the SPC receives a separate coded input for each complete sentence or message that is synthesized. This input code is received by the SPC through the SW 1–8 port.

The circuit shown in *Figure 6* uses a mechanical switch group to interface the SPC while the *Figure 7* circuit uses a hardware logic controller to input the coded message control data.



(a) Minimum Low-Pass Filter/Amplifier



(b) Maximum Filter Response Configuration

FIGURE 4. SPC Filter and Amplifier

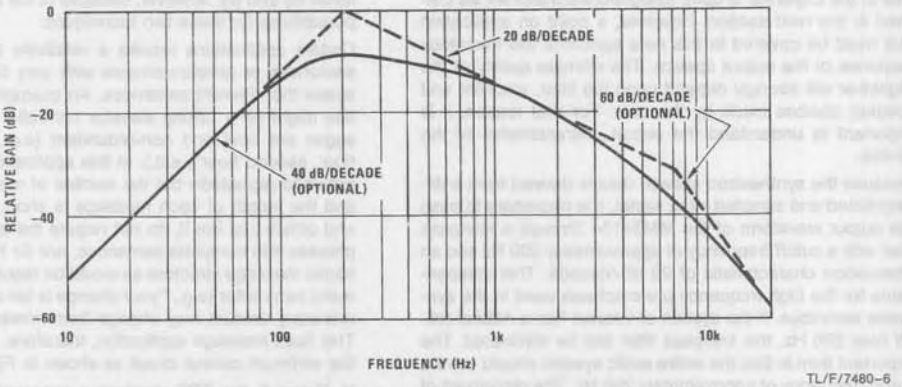


FIGURE 5. Recommended Frequency Response of Entire Audio System for MM54104 SPC

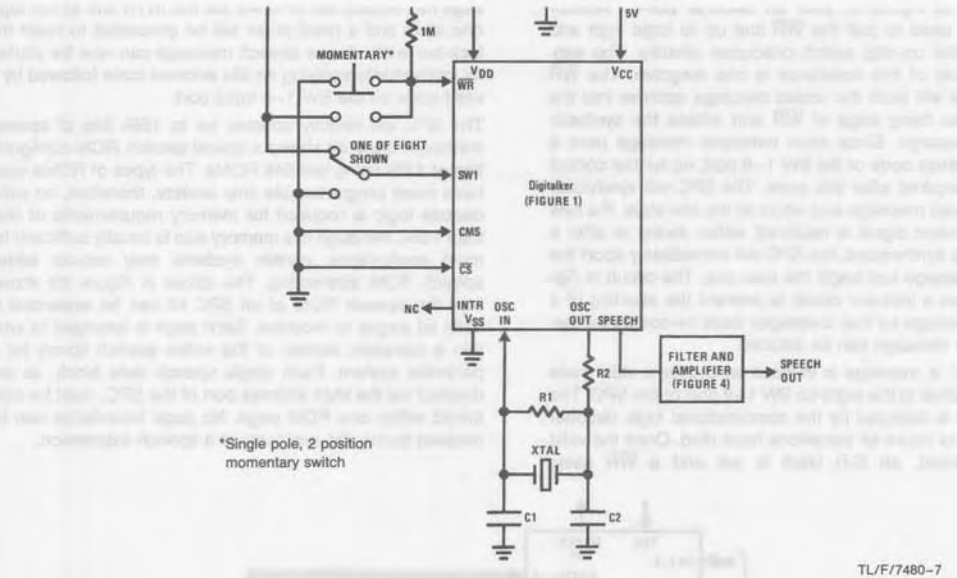


FIGURE 6. Digitalker with Switch Interface

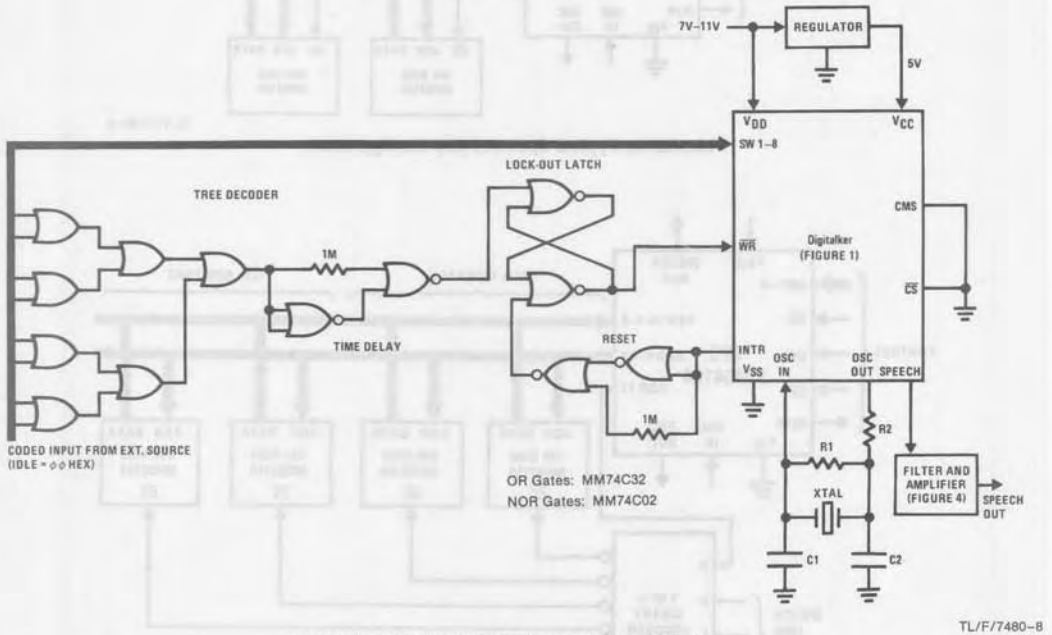
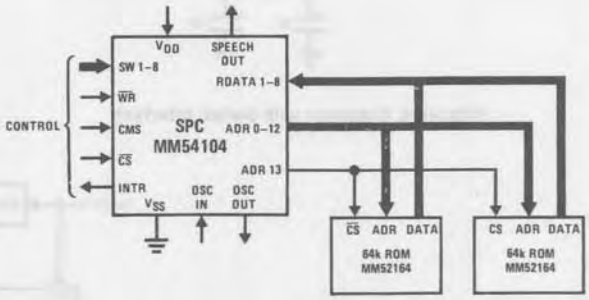


FIGURE 7. Digitalker with Logic Control Interface

After the proper message address is established on the SW 1–8 port, a momentary pulse must be applied to the $\overline{WR}$ line. If this signal is applied with a momentary action switch, as shown in *Figure 6*, then an external pull-up resistor should be used to pull the $\overline{WR}$ line up to logic high and complete the on-chip switch debounce circuitry. The suggested value of this resistance is one megohm. The $\overline{WR}$ input signal will latch the coded message address into the SPC on the rising edge of $\overline{WR}$ and initiate the synthetic speech message. Since each complete message uses a unique address code of the SW 1–8 port, no further control action is required after this point. The SPC will synthesize the requested message and return to the idle state. If a new input command signal is received, either during or after a message is synthesized, the SPC will immediately abort the current message and begin the new one. The circuit in *Figure 7* shows a lock-out circuit to prevent the aborting of a current message so that messages must be completed before a new message can be initiated.

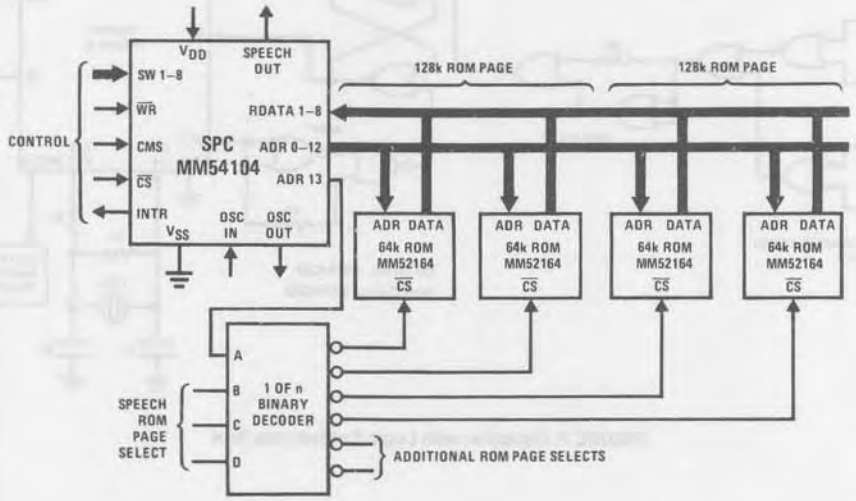
In *Figure 7*, a message is initiated whenever a valid code word is applied to the eight-bit SW 1–8 port of the SPC. The valid code is detected by the combinational logic decoder and timed to insure all transitions have died. Once the valid code is timed, an S-R latch is set and a $\overline{WR}$ rising edge is generated to start the SPC. This latch circuit also prevents retriggering of the SPC until after the present speech message is completed. Once the synthesized message has ended, the SPC will set the INTR line to the logic one state and a reset pulse will be generated to reset the lock-out latch. A new speech message can now be started by momentarily applying an idle address code followed by a valid code on the SW 1–8 input port.

The SPC will directly address up to 128k bits of speech memory. *Figure 8a* shows a typical speech ROM configuration of 128k using two 64k ROMs. The types of ROMs used have mask programmable chip selects, therefore, no extra decode logic is required for memory requirements of less than 128k. Although this memory size is usually sufficient for most applications, certain systems may require added speech ROM addressing. The circuit in *Figure 8b* shows how the speech ROM of an SPC kit can be expanded in 128k bit pages or modules. Each page is arranged to contain a complete portion of the entire speech library for a particular system. Each single speech data block, as addressed via the start address port of the SPC, must be contained within one ROM page. No page boundaries can be crossed during the synthesis of a speech expression.



TL/F/7480–9

FIGURE 8a. Typical Speech ROM Configuration



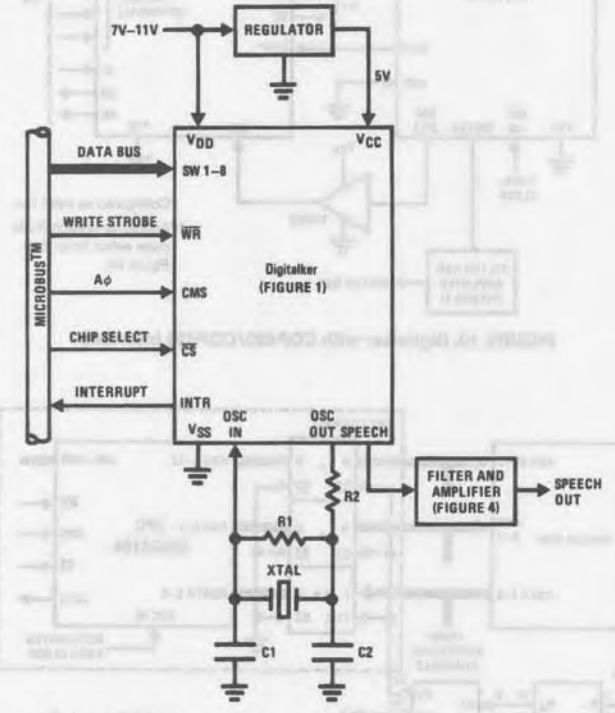
TL/F/7480–10

FIGURE 8b. Speech ROM Expansion Configuration

While the simple control schemes discussed so far can be used in many applications, a far more important group of applications will take advantage of the SPC's ability to construct sentences from a group of words, sounds and phrases. This type of application uses an intelligent controller or a microprocessor to string together a group of synthesized phrases to form a complete sentence. The electronic bank teller, previously mentioned, is a good example of this application. The microprocessor controls the stringing of SPC code addresses and applies them, one at a time, to the SW 1–8 port of the SPC. Handshake timing between the microprocessor and the SPC is provided with the INTR line. This microprocessor interface arrangement is known as Microbus and the configuration is shown in *Figure 9*.

The use of a microprocessor controller expands the versatility of the SPC tremendously. Messages that are composed of numerical responses or fixed phrases in random sequence can be easily constructed from a library speech memory. In addition, various tones or warnings can be synthesized and added before, during, or after an announcement to identify the urgency of each message. For example, an automobile message may state that "oil pressure is low". Alone, that message may only mean that pressure has dropped but no immediate hazard exists. If, however, pressure has dropped below a critical value, the message could be compounded to say "warning, oil pressure is low, pull over and stop the engine". In this latter case, phrases of high urgency are added to the initial message to increase its level of importance. Of course, the second message is not completely separate from the first but is, instead, an expansion of the first. This technique allows fewer input address codes to initiate a larger number of messages without assigning a separate address code for each message and for each of its derivatives. This would be particularly important to an electronic bank teller since a large number of monetary amounts must be synthesized for a relatively small number of finished sentences.

When preparing a speech ROM for an SPC that will synthesize whole sentences from groups of phrases, it is important to note the desired inflections. The SPC has the ability to synthesize all of the important speech attributes including pitch and gain variations, emphasis, inflection, etc. This leads to very high quality life-like synthetic speech if the stringing of phrases does not result in an artificial emphasis or inflection. It is important to choose phrases carefully and to record them with the attribute required for a realistic sentence string. The stringing of phonemes should be avoided whenever possible because the natural inflection is usually lost in such an arrangement.



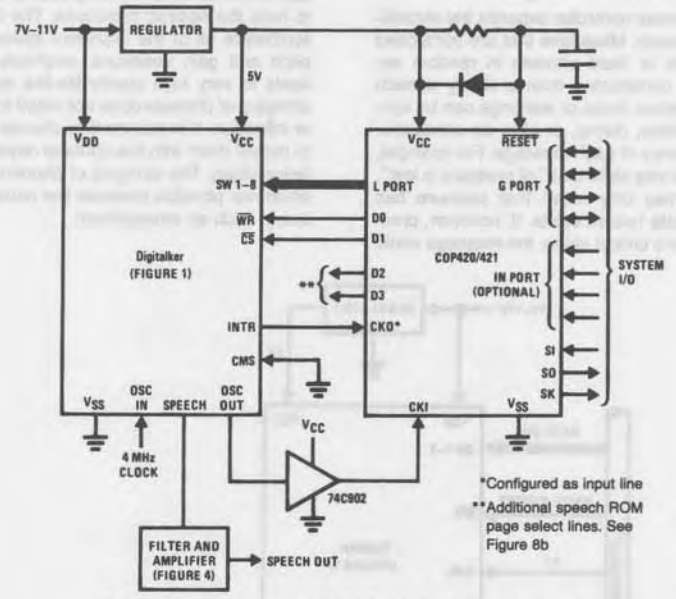
TL/F/7480–11

FIGURE 9. Digitalker with Microprocessor (Microbus) Interface

A low cost intelligent controller for the SPC is one of the COP400 series of microcontrollers. *Figure 10* shows one possible arrangement of an SPC system and a COP420. The COP provides all of the advantages associated with a Microbus interface at a relatively low cost. Because of its limited I/O structure, the COP's serial I/O port is expanded as required to obtain the desired number of input lines.

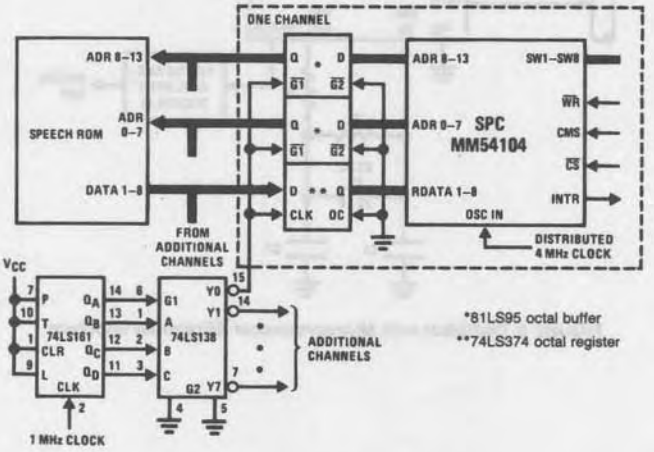
The final application technique to be covered is the multichannel configuration. The previous arrangements used an SPC and a dedicated set of speech ROMs to provide a single channel of synthetic speech. Appliances, autos, toys and games, terminals, etc. would probably use a single channel SPC arrangement. But an entirely different group of products could take advantage of a multiple channel approach to reduce the ROM requirements. This group of products includes multiple elevator controllers, electronic bank tellers, multiple pupil learning centers, voice response telephone answering equipment, telephone switching system call announcement centers, etc. In this application, each channel would use a separate SPC and amplifier circuit, but several channels would share a common controller and speech library ROM. A typical configuration is shown in *Figure 11*.

The library ROM of *Figure 11* is shared over eight SPC channels. Each SPC channel is scanned once in 16 μs as



TL/F/7480–12

FIGURE 10. Digitalker with COP420/COP421 Interface



TL/F/7480–13

FIGURE 11. Multichannel Speech Synthesizer

shown in *Figure 12*. During each channel period of 2 μs, an SPC output address is presented to the ROM address input port via a pair of octal TRI-STATE® bus drivers. After one μs, the data from the ROM is clocked into the channel's octal data latch, the output of which is connected to the SPC ROM data input port. The remaining 1 μs of each channel cycle is provided for bus settling time.

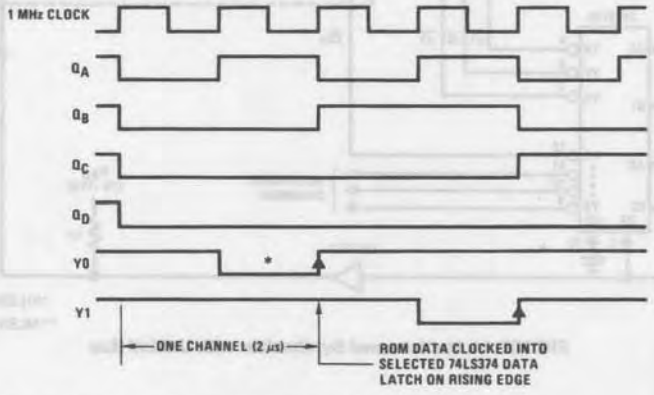
When the speech library ROM is shared over many channels, the actual number of shared channels is controlled by the MM54104 SPC memory cycle timing. Because the channel scanning is asynchronous to the SPC cycle timing, it is necessary for each channel to be scanned at least once during the high interval of the $\overline{\text{ROMEN}}$ signal. As shown in *Figure 13*, this signal is high for at least 20 μs of each memory fetch cycle. Thus, a scanning rate of one channel every 16 μs will insure that each channel is scanned at least once while the $\overline{\text{ROMEN}}$ signal is high.

One final note is necessary about the configuration in *Figure 11*. Simple modifications to the counter and decoder circuitry would allow this circuit to handle sixteen channels. A four-line to sixteen-line decoder would replace the three to eight decoder and the clock would directly enable the decoder during the logic low clock period. All sixteen channels would be scanned every 16 μs and the scan interval for each channel would be one μs—one-half μs of memory access time and one-half μs of bus guard time.

The last multichannel circuit is shown in *Figure 14*. This scheme reduces the number of wires needed between the speech ROM and each SPC channel. By multiplexing address and data over the same parallel bus, fewer wires are needed. This approach is particularly attractive when each SPC channel is located on an individual circuit card. A telephone central office or PABX announcement system is a typical example of a channel per card arrangement. *Figure 14* represents that type of system.

As shown in *Figure 15*, each channel of the unified bus approach is scanned for one μs. As many as sixteen channels, therefore, can be scanned during the $\overline{\text{ROMEN}}$ high cycle of any SPC. During each channel scan, the bus is gated to transmit the ROM address to latches on the ROM circuit board. The address is sent in two bytes. After a brief delay of one-half μs, the bus is gated to return the requested ROM data to the same SPC channel. This data is then latched on the SPC channel card. This scheme is very straightforward. It exchanges reduced interconnect wiring for additional logic circuits.

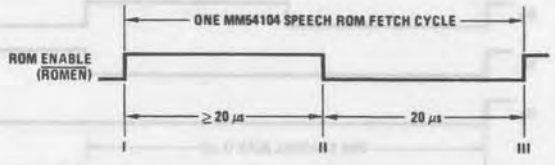
To minimize interconnect wiring when using a unified bus structure, the SPC control logic would probably be configured on a per channel basis. The COP microcontroller,



TL/F/7480–14

***Note:** Selected 81LS95 output enabled during this interval. Output is current ROM address code from SPC.

FIGURE 12. Multichannel Timing Diagram



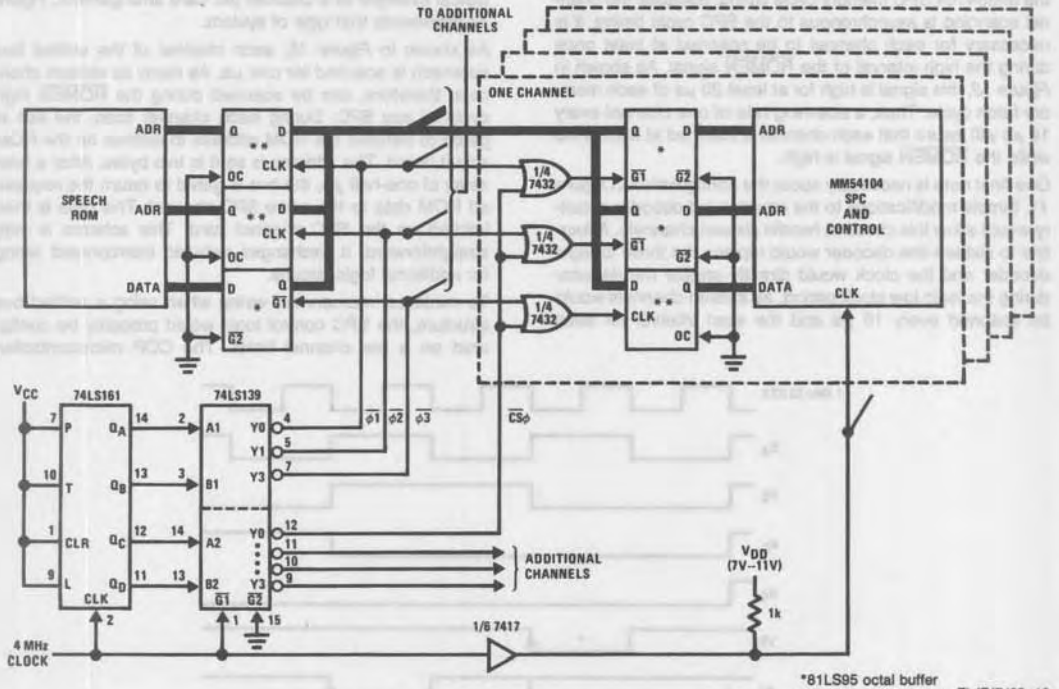
TL/F/7480–15

1. SPC about to output the next ROM address. Address will remain valid for remainder of the cycle.
2. Valid ROM data must be available to SPC shortly after this point.
3. Cycle ends, new cycle begins immediately.

FIGURE 13. MM54104 SPC Speech Memory Cycle Timing

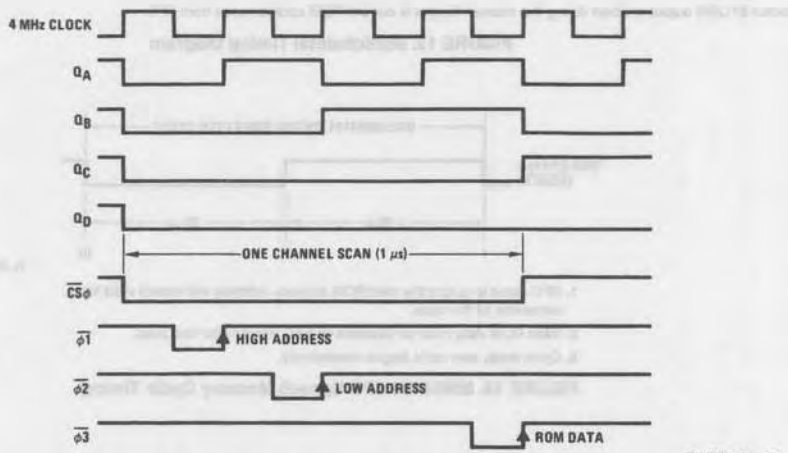
number. The COP controller reads a group of program switches or receives a down-loaded number from the switching systems's central processor. It then inserts this

the name of each decimal digit. The library could also contain the names of the teen digit pairs, and the words "hundred" and "thousand". These would be used to synthesize the words "thirteen hundred" or "two thousand", etc.



TL/F/7480–16

FIGURE 14. Multichannel Synthesizer with Unified Bus



TL/F/7480–17

FIGURE 15. Multichannel (Unified Bus) Timing Diagram

SUMMARY

This application note describes some of the versatility and flexibility of the National Semiconductor Digitalker System. This system provides low cost speech and tone synthesis for products ranging from consumer items to computers. It provides a reliable alternative to mechanical systems (i.e., tape decks) without sacrificing voice quality. Also described in this note are a few of the most popular circuit arrangements possible with the Digitalker. Of particular interest are the methods outlined to assemble whole messages from phrase groups and the schemes used for multichannel synthesizer systems. This latter application is particularly interesting because of the memory savings for the multichannel user.

REFERENCES

1. Morris, Dennis E. and Weinrich, David W., *A New Speech Synthesis Chip Set*, IEEE International Conference on Acoustics, Speech and Signal Processing, 1980.
2. Mozer, Forrest, *Method and Appartus for Speech Synthesizing*, Pending US Patent.
3. Weinrich, David W., *A Speech Synthesis Chip Set Using Compression Techniques*, Electronics, April 10, 1980.