G. Knauft
H. Lamparter
W. G. Spruth

Some New Methods for Digital Encoding of Voice Signals and for Voice Code Translation

Abstract: Recently developed techniques are described for improving the speech quality of voice signals that are first digitally encoded, placed in random access storage, and on demand are then translated into normal speech in an audio response unit under the control of a host processor. The development is an extension and modification of the channel vocoder principle. Speech quality is enhanced by hardware and software features for treatment of unvoiced components of the coded speech signal in particular by separating harmonics from the excitation function digital signal before smoothing. A new program of bit selection is used to assure that the aggregation function digital signal carries maximum information. In addition, an efficient method of storage assignment is shown for the excitation function and the aggregate function registers in the voice code translator.

1.0 Introduction

This paper reports some of the development work to improve the speech quality of the IBM 7772 audio response unit that utilizes an improved method for digital encoding of voice signals. The audio response unit permits the use of an ordinary telephone as an inquiry terminal and offers an immediate connection to the random access files of a data processing system. An inquiry is dialed or keyed from the telephone, and the answer from the data processor is received by the requester in the form of a verbal reply.

The general mode of operation of computer-controlled, voice-answer-back systems has been described in the literature.[1,2] The inquiry from the telephone-type input terminal consists of a series of digits or characters. In a typical operation the running program of the data processing system may be interrupted whenever an inquiry is completed. The input message is then assembled in core storage. The message is next interpreted by an object program that retrieves the requested data from the file. These data are then used to generate an appropriate response, which is transmitted to the audio response unit, translated into audio signals and sent to the inquirer. In the particular system described here in Section 2.0, a vocabulary is stored in "Digitally Coded Voice" (DCV) in a section of the random access file, as indicated in Fig. 1. This vocabulary is produced in a separate Voice Code Generator facility. Some of the voice code generation principles and their interplay with the audio section are discussed in this paper. A review of the conceptional features of the Integrated Vocoder that preceded the development of this system has been made by Rothauser.[3]

The DCV technique to be described here is an extension of the channel vocoder principle invented by Dudley[4] at Bell Telephone Laboratories. A large portion of the hardware has been based on a channel vocoder developed by E. Rothauser and his associates[5a,5b] at the IBM Vienna Laboratory. A channel vocoder, in reducing the quantity of information required for speech signal representation, takes advantage of the great redundancy in human speech. It breaks the speech signal down into spectrum and pitch information in an "analyzer". The manner in which this is done permits elimination of a large part of the redundancy. A "synthesizer" reproduces speech sounds from this information.

The front end of the Voice Code Generator may be regarded as a vocoder analyzer, and the Voice Code Translator (VCT) in the audio section is the equivalent of a vocoder synthesizer.

Vocoders have been recently simulated on large scale computers.[6,11,12,13] This approach produces greater flexibility and reliability but consumes a large amount of data processor time. For the Voice Code Generator a hybrid solution has been chosen, where some functions are performed by hardware and others by programmed

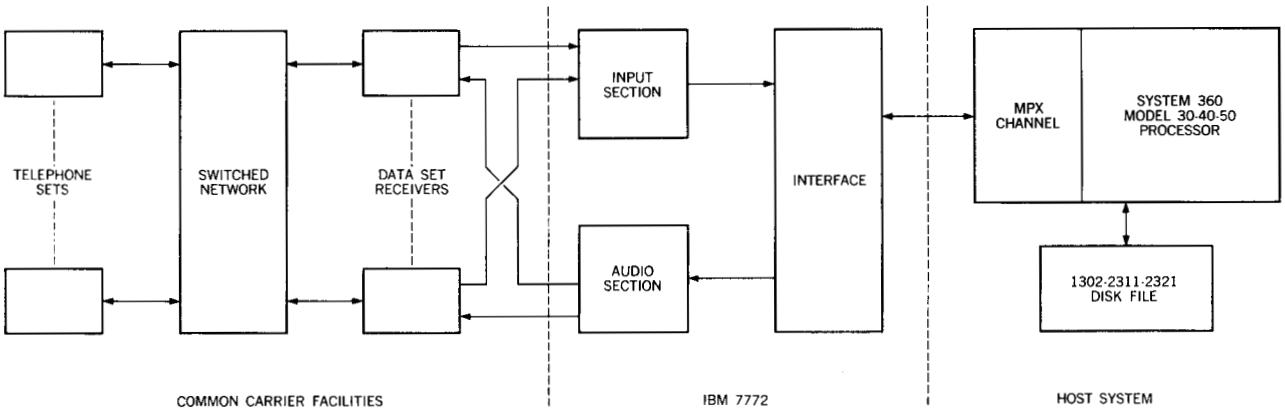


Figure 1 Audio Response System.

computer simulation, thus gaining most of the advantages of both approaches. The Voice Code Generator consists of these sections, as indicated in Fig. 2:

1) An excitation channel to measure the fundamental frequency of the speech signal.
2) Fifteen band-pass filters, covering continuously the voice spectrum from 200 cps to 3.5 kcps. Depending on the audio input, the amplitude of each band-pass filter signal is measured at intervals ranging from 3 to 20 ms; this section is called the *aggregate channel.*
3) A digital converter section, which converts the electrical signals of the excitation and aggregate channels into digital data.
4) An IBM 1401 tape system, which receives these digital data, processes them and punches them out on punched cards in the proper format. The data contained in these cards constitute the DCV equivalent of the speech signal.*

Originally, the vocoder was invented as a component for a speech signal transmission system. In this application, it has to work with a range of more and less qualified speakers, with less than optimum microphones and some environmental noise, and under real-time conditions. So far it had only limited use, mainly due to poor speech quality. Recent improvements are discussed in Refs. 7, 8, 12, and 13.

In an audio response unit application, the above restrictions do not apply. The speech quality problem can be partially solved by using a well-trained speaker and high quality recording equipment to produce the audio input signal for the voice code generator. In addition several techniques were developed to improve speech quality. These are:

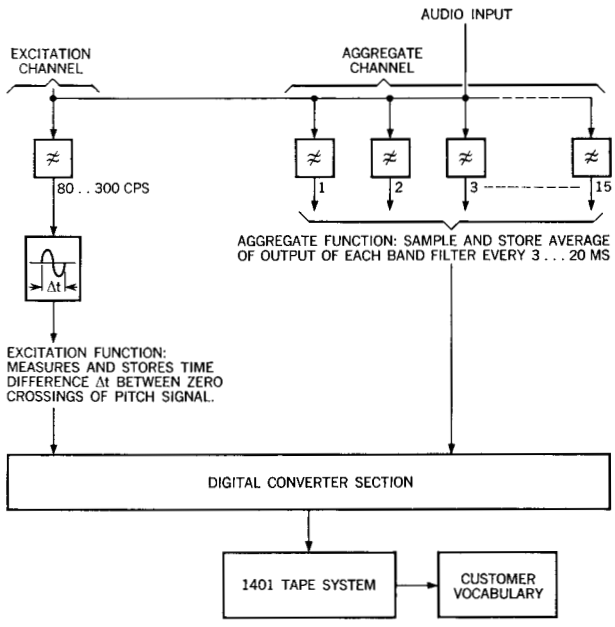
1) Treatment of unvoiced components in the speech signal. (Sec. 3.1).
2) Separating parasitic effects from the excitation function. (Sec. 3.2).
3) Variable aggregate function sampling frequency. (Sec. 3.3).

It should be noted that these techniques, while improving DCV speech quality, do not necessarily work in a vocoder transmission system.

2.0 System description

Considering the hardware only, the developmental system consists of the following functional sections (see Fig. 1):

Figure 2 Voice Code Generator.



* The IBM 1401 system included 16K core memory, advanced programming features, serial input-output adapter for attachment of the hardware described under 1) to 3), four IBM 729 tape units, 1402 card reader punch and 1403 printer.

The input section accepts the digital input message from the terminal through a data set. The data are received serial by character, translated into 8-bit code, and transmitted to the processor.

The interface control section controls the data flow between the audio response unit and the host processor through the multiplexor channel in both directions.

The audio system is comprised of two subsections. The data register implements timing requirements for the data transfer from the processor, stores from 6 to 20 ms of DCV and additional control information for multiplexing operations.

The Voice Code Translator accepts data in Digitally Coded Voice and converts it into audio signals. An average of approximately 2500 bits are required for each second of audio output.

In a workable audio response system, the following external requirements play an important role:

1) Common carrier facilities
2) Host system hardware and control programs
3) Audio response unit
4) DCV vocabulary, produced by the Voice Code Generator.

The DCV vocabulary is made available to the customer in the form of punched cards or other machine-readable documents. The customer may select from a master vocabulary list those words he wants to use in his installation. These words, which are loaded by conventional means into his mass storage, are an integral and important part of his system.

3.0 Techniques to improve DCV speech quality

The digital data sent to the 1401 system in Fig. 2 are redundant. Processing these data results in a net reduction in data rate, and a simultaneous improvement in speech quality. The techniques described in the following sections are employed.

3.1 Treatment of unvoiced components of the speech signal

3.1.1 Continuous aggregate function sampling

Voiced components of the speech signal are characterized by a line spectrum consisting of a fundamental frequency (the pitch or excitation frequency) and harmonics with variable amplitude. Such a spectrum is shown in Fig. 3. Unvoiced components of the speech signal are characterized by a continuous spectrum with little energy in the band filter used in the excitation channel. (Speech spectrums are described in Ref. 9). This band filter ranges usually from 80 . . . 300 cps. Such a spectrum is shown in Fig. 4.

To produce the DCV representation of a speech signal, the output of the aggregate and excitation channels (see Fig. 2) are sampled. A DCV recording contains

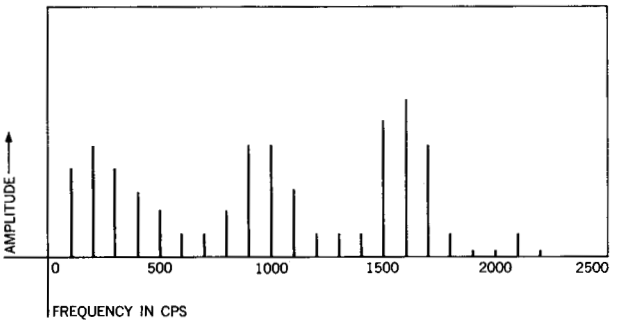


Figure 3 Voiced sound spectrum.

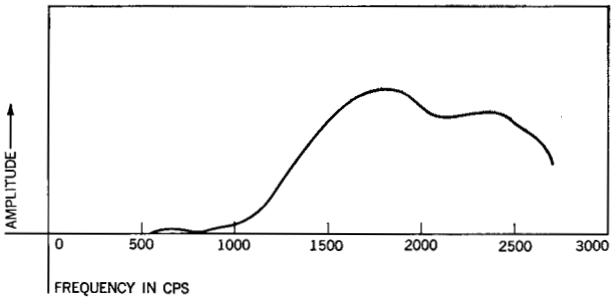


Figure 4 Unvoiced sound spectrum.

two types of interlaced digital data, which are called the Aggregate Function (AF) and Excitation Function (EF).

An AF represents the signal amplitudes of the 15 band filters in the aggregate channel section. Each one of the 15 amplitudes is coded as a 3-bit binary number permitting differentiation between 8 possible amplitude levels, 5 db apart. An AF thus requires 45 bits. With an 8-bit code, and allowing for a few control bits, each AF requires 6 characters. An EF represents the time between two adjacent aggregate channel samples. A zero-crossing detector in the excitation channel produces a pulse for every positive-going zero crossing of the fundamental frequency of the speech signal. The time Δt between two such pulses is coded as a 7-bit binary number, indicating a multiple of 200 μs. With an additional control bit, each EF requires one 8-bit character.

Each AF has an associated EF (see, however, Section 3.3). A continuous serial string of EF and AF characters represents a word in DCV. Sampling of the aggregate channel is done under control of the zero-crossing detector if there is enough energy in the excitation channel to trigger this detector. This is usually the case for voiced sounds. In particular, the aggregate channel band filters are sampled at the following points in time:

a) If the excitation channel signal is sufficiently high,

every positive-going zero crossing causes an AF sampling.

b) In the absence of a sufficiently strong excitation signal the AF band filters are sampled at 4-ms intervals.

Mode a) corresponds, roughly but not always, to a voiced speech signal segment. Mode b) is the approximate equivalent of an unvoiced signal. After each zero-crossing sample pulse, the Voice Code Generator checks an interval of 20 ms for the next zero crossing. If no pulse appears, a 4-ms counter is activated, which forces an AF sampling every 4 ms up to the point where the excitation channel signal is strong enough to produce zero-crossing pulses again.

Figure 5 illustrates this. Figure 5(a) shows the pulse sequence emitted by the zero-crossing detector, 5(b) the pulse sequence emitted by the 4-ms counter, and 5(c) the pulse sequence used for actual AF sampling.

In order to perform this operation, the excitation channel pulses originating from the zero-crossing detector (Fig. 6) are

1) used to set a 20-ms single shot serving as "interval timer" and
2) routed through a 20-ms delay and then into the IBM 1401 core memory.

The IBM 1401 processor measures the time between the arrival of adjacent zero-crossing pulses. This is done by counting the number of 1401 memory cycles with the help of a move instruction. Whenever a zero-crossing pulse is received the processor performs these operations:

1) Stores the time difference measured in multiples of 200 μs and expressed as a 7-bit binary number, as another new EF value.
2) Originates another AF sample by once stepping through the AF address counter.
3) Detects, by interrogating the 20-ms single shot, whether another zero-crossing pulse can be detected within the next 20 ms.

In the case no further zero-crossing pulse will arrive during the next 20 ms, the program branches into another subroutine, which forces an AF sampling every 4 ms. This is done up to the point in time where another zero-crossing pulse arrives. A tag bit associated with each AF defines whether the sampling was forced or caused by a natural zero-crossing pulse.

This method, while not providing a clear voiced-unvoiced decision, assures a continuous sampling of the aggregate function signal. It is not dependent on the accuracy or inaccuracy of excitation channel recordings. More conventional approaches usually link in one form or another the voiced-unvoiced decision to the AF

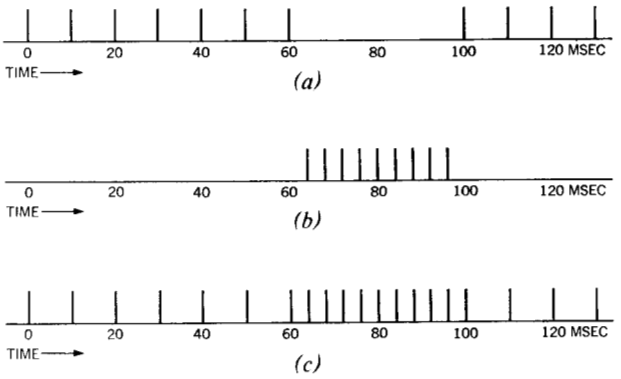
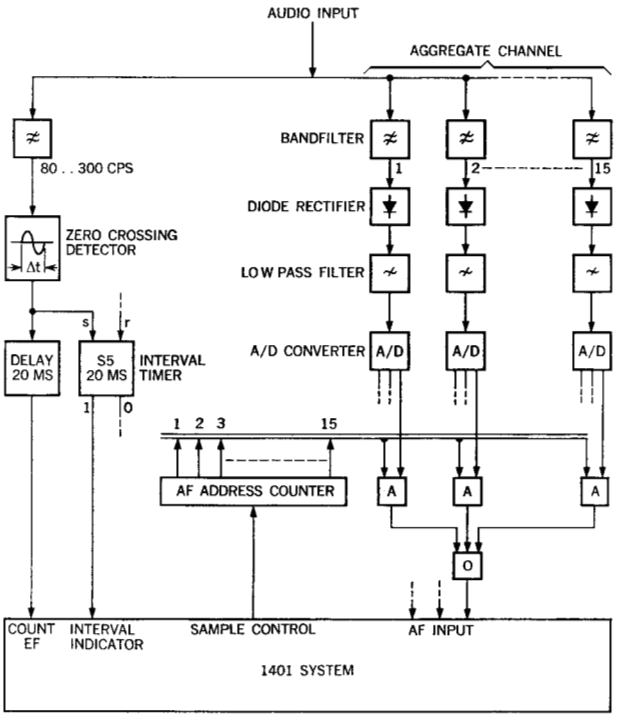


Figure 5 Sample pulse generation.

Figure 6 Voice Code Generator.



sampling. This can result in an improper EF representation or in the missing of important AF samples.

3.1.2 Reprocessing DCV data by the audio program

The voice code generator in Fig. 6 contains only a minimum amount of hardware in addition to the IBM 1401 system. The data read into the 1401 *could* be sent directly to the VCT and translated back into audio. However, speech quality of this output signal would be poor.

To improve this, data originating from the voice code

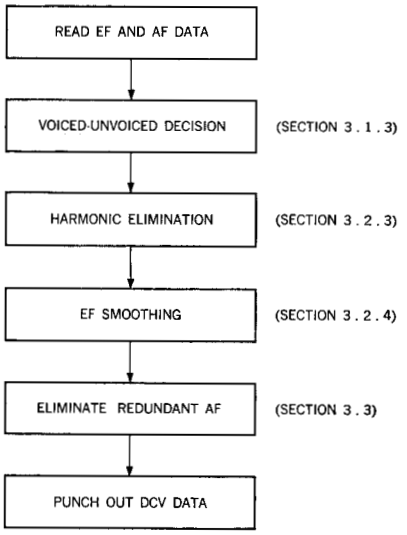


Figure 7 The audio program.

generator analog hardware have to be further processed. In conventional vocoder systems, this can be done in analog form through additional special circuits in the aggregate and excitation channels. Digitalization by PCM is the very last step in the voice analyzing process.

Other than in an audio transmission system the voice code generator does not have to produce a DCV vocabulary in real time. The digital data sent to the 1401 in Fig. 6 contain all important information. For a skilled person it is not particularly difficult to reduce or modify these digital data to obtain optimum speech quality. We therefore preferred to accept the raw data from the aggregate and excitation channels, and process them in digital form in a specially developed audio program.

The audio program consists of a control monitor and individual subprograms. The control monitor receives the input data and processes them consecutively with each one of the subprograms, as shown in Fig. 7. The function of the individual subprograms is discussed below.

Data from the aggregate channel entering the audio program have much redundancy. This eases the voiced-unvoiced decision process (Section 3.1.3). Later elimination of redundancy causes no serious problem (see Section 3.3). The harmonic elimination program described in Section 3.2.3 also reduces the bit rate.

3.1.3 Voiced-unvoiced control

The first subprogram of the audio program makes the decision as to whether a speech element is voiced or unvoiced. This decision is made for each individual EF and is based on an inspection of this EF, the preceding and following EF and the associated AF. Each AF contains a flag bit, which is set to:

ONE, if this AF receives unvoiced excitation
ZERO, if this AF receives voiced excitation.

Each EF has an associated AF. Essentially, termination of the EF time measurement triggered the sampling of the aggregate-channel band filter outputs. If this sampling was caused by a pulse emerging from the zero-crossing detector, the sample is not forced. If the AF sampling was initiated by the 1401 because interval time in Fig. 6 was reset, the sample is forced.

The part of the AF input to the 1401 originating from band filters 1, 2 and 3, representing the frequency band from 200 to 550 cps, is treated as a 3-digit octal number. A decision is made as to whether or not this number is larger than 122.

Essentially, the decision number rules are these:

1) Forced samples receive unvoiced excitation, unless filter $1 \ldots 3 \geq 122$.
2) Unforced samples receive voiced excitation, unless an EF has forced neighbors and filter $1 \ldots 3 < 122$.

A particular situation arises if a string of at least 3 consecutive EF's is not forced, but has amplitudes Filter $1 \ldots 3 < 122$. If this happens, a further test is made. The EF still receives voiced excitation. However, if the audio signal energy in the frequency band 200 . . . 550 cps is relatively weak (as seen by inspection of the AF), a special marker bit contained in the AF is set. This marker bit is later eliminated during the smoothing program (Section 3.2.4) but excludes this particular AF from participation in the smoothing process.

This case is representative for strong voiced fricatives, for example, the "g" in the French word "agent".

3.1.4 Statistical excitation in the Voice Code Translator

A detailed diagram of the VCT is shown in Fig. 8. The DCV data arriving from the host system CPU or I/O channel are separated into their AF's and EF's. The AF is loaded into an AF register and the EF into an EF counter. The band filters are excited with narrow pulses. Their amplitude is determined by the AF, and the pulse frequency—the same for all filters—by the EF. This is done by stepping the filter address counter once through its 15 positions and reading the corresponding octal digit from the AF register into the D/A converter. The data in the EF counter are continuously counted down to zero with a 5 kcps count-down frequency. Whenever a count down reaches zero, the EF counter is immediately reloaded with the next EF, waiting in the EF register.

During voiced DCV segments, the filters are excited only whenever a EF zero count-down is completed.

This time interval corresponds to the time between two adjacent positive zero crossings of the pitch signal, as shown in Fig. 2 and Fig. 6. The EF thus determines the time difference between two adjacent excitations of the output filters.

For unvoiced DCV segments, the flag bit resets the flip-flop, and thus opens the adjacent AND gate. A random sequence of pulses produced by the random pulse generator provides additional excitations for the output filters.

This operation is indicated in Fig. 9. In Fig. 9a a series of EF count-down pulses is shown. A series of pulses originating from the random pulse generator is indicated in 9b, and the pulse sequence addressing the filter address counter, if all parts are voiced, is shown in 9c. In 9d the EF pulses 2 and 3 have an associated flag bit in their AF, and require statistical excitation. During this time the pulses are the sum of 9a and 9b.

We thus may state that:

a) During voiced DCV segments, the EF determines the time interval between output filter excitations.
b) During unvoiced DCV segments, the EF determines the time duration, during which the output filters are excited with a random sequence of rectangular pulses.

The EF counter in Fig. 8 is buffered by a second EF register to assure the immediate availability of a new EF at the moment of the zero count down. Similarly the AF register has to be buffered by a second register, which may be part of the host system's core memory.

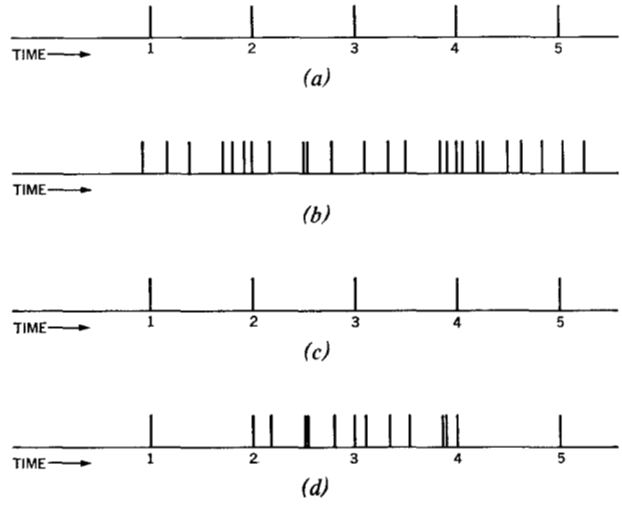
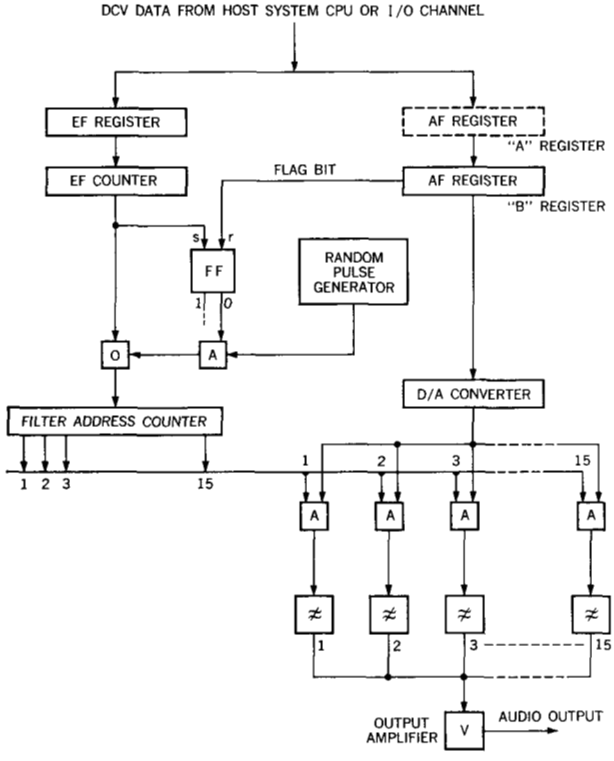


Figure 9 VCT excitation pulse sequence.

Figure 8 Voice Code Translator.



3.1.5 Statistical Excitation Frequency

If the unvoiced DCV segments are to produce an audio output of the same loudness as the voiced segments, the average pulse repetition rate for excitation of the output filters has to be the same for voiced and unvoiced segments. Listener tests, and a theoretical analysis, indicate that for faithful reproduction of unvoiced sounds the random pulse generator in Fig. 8 has to have a mean pulse repetition rate several times higher than the pulse repetition rate used for voiced sounds. This is also indicated in Fig. 9.

As a consequence, unvoiced sounds are reproduced louder than voiced sounds. To correct for this the unvoiced sounds have to be de-emphasized. This can be done by reducing the amplitude or width of the pulses exciting the output filters, or by reducing the amplification factor of the output amplifier. In the audio response unit the pulse width is reduced.

• *3.2 Separating parasitic effects from the excitation function*

3.2.1 Nature of distortions

The signal produced by the zero-crossing detector in Fig. 6 is a series of rectangular pulses, the frequency of which is supposed to be identical with the pitch frequency. For a constant pitch all pulses are supposed to have the same separation in time.

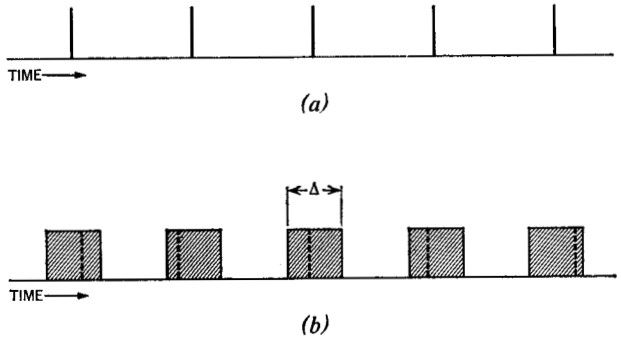


Figure 10 Small displacement of sample pulses. Actual pulse position is somewhere in hatched area as indicated by dashed lines.

Small periodicity deviations cause a noticeable speech quality distortion. This means the signal may have the form of Fig. 10b instead of Fig. 10a. The small rectangular pulses of Fig. 10a are randomly positioned somewhere in the hatched areas of Fig. 10b.

The signal of Figure 10a has the energy density spectrum $\Phi(\omega)$ shown in Figure 11a, assuming the pulse width is very small.* The energy density spectrum of the Figure 10b signal is the sum of the two spectra shown in Figs. 11b and 11c.

Figure 11b shows a line spectrum just like the one in Figure 11a. However, the amplitudes of the spectral lines fall off with higher frequency and reach the value zero at the frequency $f = 1/\Delta$. The envelope f_e of the individual frequencies follows the characteristic

$$f_e = \left(\frac{\sin \omega\Delta/2}{\omega\Delta/2}\right)^2.$$

Figure 11c shows a continuous spectrum, the amplitude of which reaches a maximum at the frequency $f = 1/\Delta$.

The periodicity deviation width in Fig. 10b re-appears in Fig. 11b and 11c as the frequency at which the line spectrum components are zero, and the continuous spectrum reaches a maximum. From these Figures it can be seen that Δ produces strong distortions of the original spectrum.

Listener tests demonstrate that a periodicity deviation of $\Delta = 100\ \mu s$ and $1/\Delta = 10$ kcps causes a noticeable degradation in speech quality.

A slight displacement of the excitation pulse may be due to random noise in the excitation channel, or an inaccurate operation of the zero-crossing detector or a change in the harmonic content of the pitch signal. It has been experimentally demonstrated that the typical "rough characteristic" of speech output from a vocoder of this type is at least partly caused by these small periodicity deviations in the pitch signal.

* Calculation of power spectrums of random signals is described, e.g., in Ref. 2.

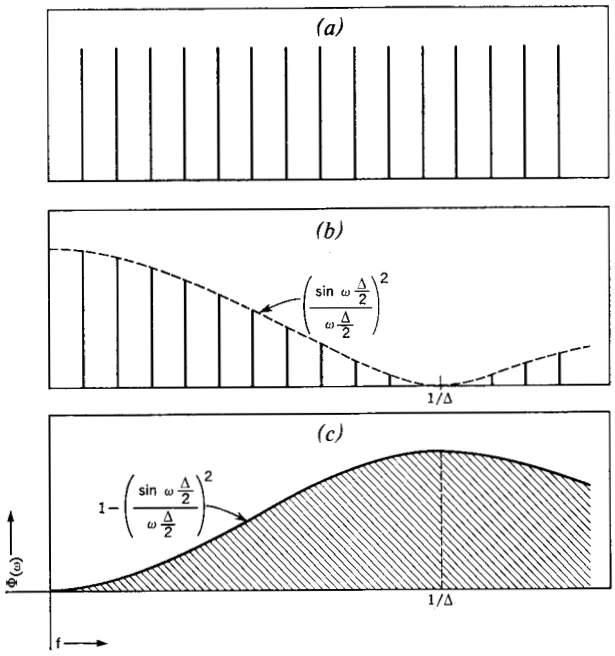


Figure 11 Sample pulse spectra.

Figure 12 Actual and ideal recording of Excitation Function.

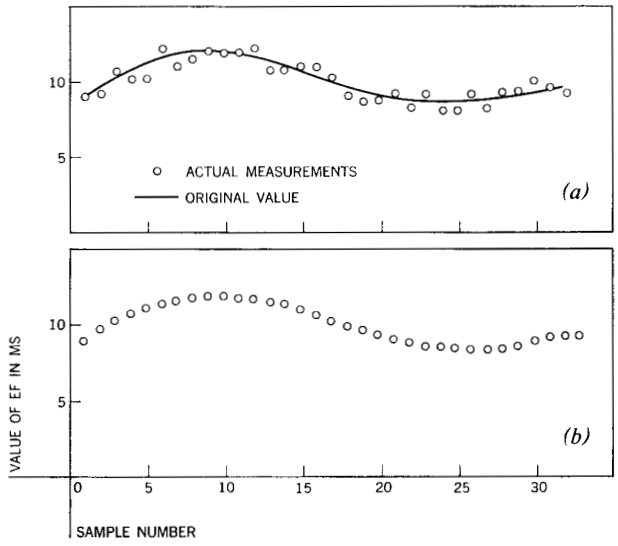


Figure 12a shows an EF signal as it originally was, and as it will be measured with small periodicity deviations. By an averaging process over several EF, we may achieve the EF signal shown in Fig. 12b, which indeed closely restores the original EF values, and thus fidelity, to the speech signal.

This procedure does not work because of the effect of harmonics. For an illustration, see Fig. 13. A growth in the effect of a harmonic causes sudden additional excitation pulses to appear.

As a result, actual measurements taken will look more like Fig. 14 instead of Fig. 12a. In this example, 8 of the measured EF's deviate very considerably from the original value. Figure 14 is very typical for actual recordings. Any averaging process obviously would worsen the situation. To solve this problem, the harmonics are removed from the EF signal before smoothing. The techniques described in the following sections are employed.

3.2.2 Low-pass filter in the excitation channel

The excitation channel band-pass filter in Fig. 6 is replaced by a low-pass filter with the following characteristic. It is flat up to 80 cps (the lowest male pitch frequency), and then falls off approximately 6 db per octave. This greatly reduces the number of additional zero crossings produced by harmonics. If the second harmonic amplitude is smaller than the fundamental, additional zero crossings will be completely and always eliminated.

For vocabulary recordings with a female speaker, a higher low-pass cut-off frequency is preferable.

3.2.3 Harmonic elimination by programming techniques

For a further elimination of harmonics, all EF values accepted by the IBM 1401 are investigated by the "harmonic elimination" program (see Fig. 7). Two passes of the EF data are required.

In the first pass, each EF is tested for the condition $EF < d$. The exact value of d depends on characteristic properties of the speakers voice (e.g., male or female) and is adjusted from one recording to another.

In the first pass, all EF's with a value smaller than d are eliminated. If the program encounters such an EF, it adds the value of this EF either to the preceding or the following EF, whichever is smaller. The sum replaces the two individual EF's.

The first program pass takes advantage of the fact that the pitch frequency of each speaker has a maximum. If a particular EF exceeds this limit, we are sure that we have recorded a harmonic. In a planned improvement of this first pass the program itself will determine the proper value of the constant d.

In the second pass of the harmonic elimination program, most of those EF's are eliminated, which do not exceed d, but still represent a harmonic. This is done by comparing the sum of two adjacent EF values, called EF_n and EF_{n+1}, against both the preceding value EF_{n-1} and the next following value EF_{n+2}. If the sum falls within 1 ms of either EF_{n-1} or EF_{n+2}, the pair of EF's is eliminated and replaced by a single new EF with the value $(EF_n + EF_{n+1})$.

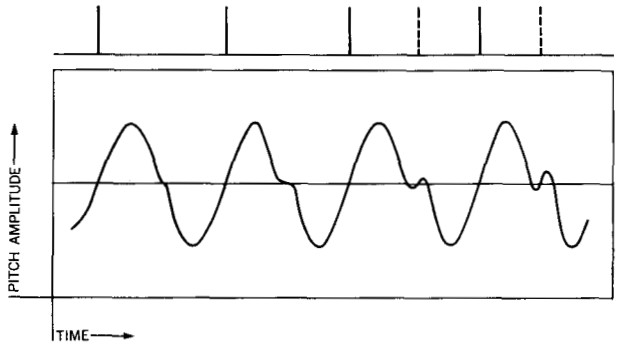
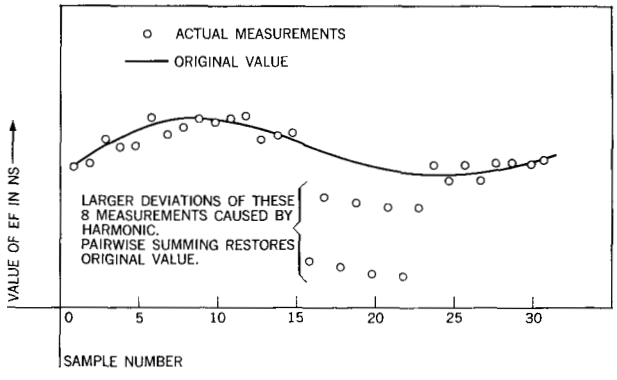


Figure 13 Harmonics in excitation channel. Solid-line pulses are due to fundamental; dashed-line to harmonics.

Figure 14 Recording of Excitation Function with harmonics.



The actual harmonic elimination program contains additional refinements, in particular to handle starting and ending procedures. In connection with filter described in Section 3.2.2, nearly all the harmonics are eliminated.

3.2.4 Smoothing the Excitation Function

The smoothing program processes all voiced EF's, after they have passed through the harmonic elimination program. A weighted average of 11 consecutive EF's is formed. The weight is greatest for the center EF, and decreases linearily at both sides. The weighted average replaces the value of the center EF. This process is repeated for all EF's in a given DCV word. For final elimination of harmonics, all EF's which deviate by more than 1 ms from both their neighbors are excluded from the average forming process. The smoothing process also stops at sharp transitions from one EF to the next, as they sometimes occur at voiced stops in a word.

AF's which have the marker bit mentioned in Section 3.1.3, are representative for strong voiced fricatives. They are excluded from the smoothing program, and retain their original EF values. (The marker bit is stripped at this point.) This results in an excitation, which is partly voiced, partly unvoiced. The spectrum of the

excitation signal is the sum of the two spectra shown in Figs. 11b and 11c. This mixed excitation produces better reproduction of voiced fricatives than either voiced or unvoiced excitation. In such cases, where the original sound wave contains voiced as well as unvoiced components, smoothing of the excitation function produces negative results.

• *3.3 Variable Aggregate Function sampling frequency*

The speech synthesizer used by the voice output system requires the value of the aggregate function to be always concurrent with an excitation pulse. Thus the analyzer should sample the aggregate function whenever an excitation pulse has been detected. This sampling technique, however, would produce a very high information rate. The following techniques could be used to reduce the information rate:

1) The AF is sampled only every n excitation functions, where n has a fixed value, e.g., $n = 2, 3, \cdots 6$.
2) The AF is sampled exactly every X milliseconds, where X is a fixed time, e.g., $X = 25$ ms, and associated with the nearest EF.
3) The AF is sampled together with the first EF appearing after a fixed time has elapsed since the last AF sample. This minimum time between AF samples may have a typical value of 20 ms.

All these methods are often less than optimum because either one of two situations may arise:

1) AF data are redundant because consecutive AF's are identical or nearly identical. This appears during periods where the AF changes very slowly.
2) In fast-changing segments of a word, important AF data are missing. This happens particularly at unvoiced-voiced transitions.

To improve this, a fairly large number of AF samples is stored in the IBM 1401 memory, one for each zero crossing pulse and one every 4 ms during the absence of a strong signal in the excitation channel. Afterwards, a special program eliminates the majority of AF samples. The remaining AF's carry a maximum of information.

The AF elimination program works in the following way. A particular aggregrate function AF is compared against its predecessor AF_{n-1}. If the difference $AF_n - AF_{n-1}$ does not exceed a preset level Δ, the AF_n is eliminated, and AF_{n+1} is compared against AF_{n-1} in exactly the same way. If $AF_n - AF_{n-1} > \Delta$ then AF_n is retained in the DCV vocabulary, and AF_{n+1} is compared against AF_n.

Thus during word segments with fast changing AF's every AF will enter the DCV word. During times with small changes, only those AF's become a part of the DCV word, which carry enough additional information to justify this.

• *3.4 Comparison of methods*

The procedure outlined above produces a DCV vocabulary with a speech quality level that is satisfactory for practical applications in a commercial environment. Some of the difficulties encountered, shortcomings, and alternative approaches are discussed below.

The audio program was difficult to produce, particularly in extreme cases of harmonic elimination. Here we suggest having an additional input to the 1401 from a throat microphone. Use of a nonlinear network in the excitation channel produced negative results. We did not experiment with a frequency tracking network and cannot report on this. In some cases the treatment of voiced fricatives leaves room for improvement.

The choice of band-pass filters in the aggregate channel was arbitrary. It is possible that a different set of band-pass filters, both in terms of filter characteristic and band width allocation may produce superior speech quality.

A promising method for very good speech quality is the Voice Excited Vocoder proposed by E. David and M. Schroeder. We did not consider it because it required a relatively high bit rate.

We found our approach to be a very convenient concept allowing for easy and quick modifications of our DCV generating equipment. Some of the functions performed by the audio program may be difficult to implement through analog hardware.

4.0 Output channel register

As indicated in Section 3.2.1, the EF has rather tight tolerance requirements. It is important to note that this tolerance requirement applies only to individual EF's with relation to its neighbours. If all the EF's shown in Fig. 12 are uniformly displaced by a constant amount, for example ± 1 ms, speech quality will not be negatively influenced.

Figure 15 Step function representation of Excitation Function.

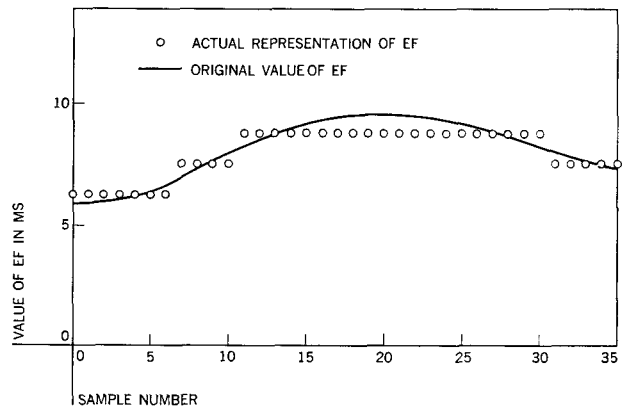
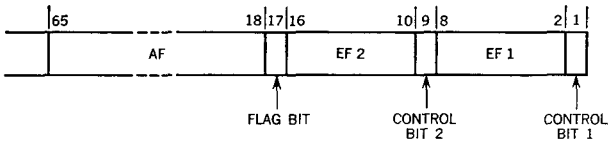


Figure 16 Storage assignment in Audio Section Register.



Moreover, it is possible to replace the smooth excitation of any particular word by a step function, as indicated in Fig. 15. Provided the steps have a difference of no more than 200 μs the effect is hardly noticeable. The original value of the excitation function may be converted into such a step function by rounding off all EF's to the nearest 200 μs.

Although this will not contribute to speech quality, it allows a very attractive hardware solution for the EF and AF registers used in the VCT.

A single delay line is used to perform the AF and EF counter and register functions shown in Fig. 8. This delay line may be 200 μs long.

The storage assignment is shown in Fig. 16 as follows:

bit position	1	control bit for first EF
	2 . . . 8	first EF (7 bit)
	9	control bit for second EF
	10 . . . 16	second EF (7 bit)
	17	flag bit (see Section 3.1.3)
	18 . . . 65	AF (48 bit, including control bits).

The data shown in Fig. 16 appear at the delay line sense amplifier in ascending order, bit 1 first, and bit 65 last. One of the two EF fields is routed through a "minus 1" network, which decreases the content of the field by 1 at each revolution. The other EF field and all remaining data bypass the "minus 1" network. Which one of the two EF fields will be counted down is determined by a 1 bit in the control bit position preceding the EF field.

At the particular delay line revolution where the count down reaches zero, a zero count-down carry bit detector will be activated. The data of the following AF field are routed to the D/A converter, 3 bits at any time. With the help of the filter address counter, which in turn is advanced every 3 bit times, the corresponding narrow pulse is used to excite the appropriate band filter.

When a zero count-down has been reached, the preceding control bit is reset to ZERO, and the other control bit set to ONE. A new count-down cycle may start. At the same time a new EF is requested from the host system. The EF enters the 8-bit interface data register, where it is serialized and read into the new empty EF field.

The total storage requirement in the delay line amounts to 65 bits. A 200 μs delay with 200-bit storage capacity allows time-sharing the delay line and most of the control logic for several VCT's. The 65-bit field for the 2nd VCT follows the 65-bit field for the 1st VCT.

5.0 Conclusion

The vocoder concept was originally conceived for use in voice transmission systems. The idea of using it in an audio response unit is new, and was proposed, among others, by the scientists at the IBM Vienna Laboratory.

We believe the combination of an unsophisticated analog voice analyzer with a powerful digital program to reconstruct original voice data and improve speech quality is new. It is a flexible and economical way to create the vocabulary for our audio response unit application. Because of the real-time factor this concept is not readily applicable for a speech transmission system.

Acknowledgments

The work described here was a joint effort of the IBM Laboratories in Vienna, Boeblingen (Germany) and La Gaude (France). The authors wish to express their thanks for contribution support, and advice received in particular from K. Bandat, E. Rothauser, H. Zemanek at the Vienna Laboratory, L. Bergmann, T. Einsele, E. Goldbach, L. Reichl at the Boeblingen Laboratory, R. Buron, E. Paris at the La Gaude Laboratory, and W. Chapman, J. Lyon, H. Wild, and W. Wolensky at the IBM laboratories in Poughkeepsie and Kingston, N. Y. A considerable amount of progress has been made at the La Gaude Laboratory since this manuscript was written. Hopefully, it will be reported in a later paper.

References

1. G. E. DuBois, "IBM Audio Response System," *Modern Communications* **3**, 10 (July–September, 1964).
2. G. E. DuBois, "Audio Response for the New York Stock Exchange," *IEEE Convention Record*, (1965).
3. E. Rothauser, "The Integrated Vocoder," (to be published).
4. H. W. Dudley, "The Automatic Synthesis of Speech," *Proc. Natl. Acad. Sci.* **25**, (1939).
5a. Ch. Schwiedernoch, "Entwicklung eines Vocoders," Diss. T. H. Wien, 1965.
5b. E. Rothauser, "Ein Impusverfahren zur Sprachübertragung," Diss. T. H. Wien, 1960.
6. L. G. Kersta, "Digital Computer Synthesizes Human Speech," *Bell Lab. Rec.* **40**, 424 (1962).
7. E. E. David, M. R. Schroeder, B. F. Logan and A. J. Prestigiacomo, "Voice Excited Vocoders for Practical Bandwidth Reduction," *IRE Trans. on Information Theory* **IT-8**, 101–105 (September 1962).

8. Paul G. Edwards, "Better Vocoders are Coming," *IEEE Spectrum* **9**, 119–129 (1964).
9. G. Fant, *Acoustic Theory of Speech Production*, Monton and Co., the Hague, 1960.
10. S. J. Mason and H. J. Zimmerman, *Electronic Circuits, Signals, and Systems*, John Wiley and Sons, Inc., 1960.
11. R. M. Golden, "Digital Computer Simulation of a Sampled-Data Voice-Excited Vocoder," *J. Acoust. Soc. Am.* **35**, 1358 (1963).
12. B. Gold, "Experiment with Speechlike Phase in a Spectrally Flattened Pitch-Excited Channel Vocoder," *J. Acoust. Soc. Am.* **36**, 1892 (1964).
13. B. Gold, "Techniques for Speech Bandwidth Compression, Using Combinations of Channel Vocoders and Formant Vocoders," *J. Acoust. Soc. Amer.* **38**, 2 (1965).

Received May 7, 1965.