Technical Note

1978-43

A Study of Future Directions for Low Rate Speech Processor Research, Development, and Implementation

B. Gold
J. I. Raffel
T. Bially

15 November 1978

Prepared for the Defense Advanced Research Projects Agency
under Electronic Systems Division Contract F19628-78-C-0002 by

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-78-C-0002 (ARPA Order 2006).
This report may be reproduced to satisfy needs of U.S. Government agencies.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Raymond L. Loiselle

Raymond L. Loiselle, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

A STUDY OF FUTURE DIRECTIONS FOR LOW RATE SPEECH PROCESSOR RESEARCH, DEVELOPMENT, AND IMPLEMENTATION

B. GOLD
J. I. RAFFEL
T. BIALLY

Group 24

TECHNICAL NOTE 1978-43

15 NOVEMBER 1978

LEXINGTON MASSACHUSETTS

ABSTRACT

This report summarizes the findings of a study which was undertaken to identify promising areas for future research and development in the narrowband voice terminal area. The objective of the study was to relate device technology developments to voice terminal concepts and designs, in order to better understand the steps that would eventually lead to very small and inexpensive narrowband terminal implementations. The overall conclusion derived from this effort can be summarized as follows: Given the present and near future advances in technology it is predictable that present-day vocoder algorithms will soon be implementable as low cost compact devices. How soon this comes to pass does not depend on any new technological breakthroughs but rather on the direction of investments in LSI development realizable with present device technology. Future systems will require terminal capabilities that exceed those of current designs, and these will result in commensurately more complex hardware implementations.

CONTENTS

I. INTRODUCTION

For a wide variety of well-documented reasons,[1] the Armed Forces have a strong, continuing interest in the development of reliable, environmentally robust, low cost yet sophisticated speech terminals. A major component of such a terminal is the speech processor, or vocoder. This report summarizes an ARPA-sponsored study undertaken by MIT Lincoln Laboratory on the subject of speech terminals. Major emphasis in this study was on the potential for substantial cost reduction of speech processors created by integrated circuit technology. To this end, various industrial and university specialists working in the field of speech processing were invited to present their ideas and summarize their own efforts in this area. Attention was directed exclusively to narrowband (less than 5 kilobits/second) processors. Summary reports of these meetings are presented in the Appendix.

The overall conclusions derived from this effort can be stated succinctly: Given the present and near future advances in technology, it is predictable that present-day vocoder algorithms will soon be implementable as low cost compact devices. How soon this comes to pass does not depend on any new technological breakthroughs but rather on the direction of investments in LSI development realizable with present device technology. However, future systems will require terminal capabilities that exceed those of current designs, and these will result in commensurately more complex hardware implementations. Thus, attention in this report is focussed on the more long range issues of improved designs.

The report focusses on long range issues of speech processor and voice terminal design, and on future expectations for LSI technology.

Included is a review of several industrial, academic and government views of the subject, a review of present day technology and its application to practical voice algorithm implementation, and an overview of the systems issues related to using narrowband speech algorithms for secure digital voice communications. We also speculate on the capabilities of future system voice terminals and predictions of what further speech algorithm research coupled with LSI development might yield in the long run.

II. DISCUSSIONS WITH INDUSTRIAL, ACADEMIC AND GOVERNMENT PEOPLE

As a preliminary activity, the subject material of this report was discussed with industrial, academic and government colleagues who were engaged in related efforts. Summary reports of these discussions are presented in the Appendix. Participants were roughly divided into two main classes, one class representing device technology development as applied to speech processors and the other class representing research workers oriented towards new speech algorithms. A variety of viewpoints were advanced which we summarize here. By and large, the university people stressed the importance of continuing research directed towards improved speech algorithms, the ultimate goal being the development of systems that are of comparable quality and robustness to conventional telephones. The device technology people were each pursuing a specific approach towards future LSI implementations, ranging from analog and digital CCD's to SOS, MOS and TTL. A delicate and still unresolved issue revolves around the microprocessor concept. A general purpose approach to speech algorithm development was advocated by some on the grounds that device technology would inevitably produce small enough LSI packages to make any future general purpose speech processor implementation very cheap. Given that, future algorithm improvements can be made with very minor design changes, such as modifying a read-only memory. On the other hand, certain attractive devices such as analog CCD's are inherently special purpose, and indeed the most promising low cost vocoder currently being designed is one employing analog CCD's.

III. PRESENT STATE OF SPEECH ALGORITHMS

Speech digitization algorithms come in a great variety of structures and at data rates varying from 600 bps to 64 kbps. The widely used continuously

variable slope delta modulation (CVSD) algorithm (at 16 kbps and 32 kbps) has been implemented as a single LSI chip. In this report we restrict ourselves to the more complex narrowband algorithms which we define as algorithms that run at lower than 5 kbps. Specifically, we review the status, as we see it, of the three major vocoder algorithms, namely, LPC (linear predictive coding), the channel vocoder and the homomorphic vocoder. Emphasis here will be placed on the structure of these algorithms with the goal of understanding requirements for appropriate LSI chips. It will also be useful to discuss the speech processing capabilities of these algorithms to serve as a reference for the more advanced ideas presented in Section V.

A. Linear Predictive Coding

Acceptance of LPC has followed the same pattern as that of other vocoder algorithms in the past. When the original work was first presented in the early 1970's, the results of a limited number of simulation runs were that the processed speech, even at rates as low as 2400 bps, appeared to compare favorably with the original speech. As a result, a feeling evolved in the US that LPC really led to a substantial improvement over previous vocoder algorithms. This improvement has unfortunately not materialized. Narrowband Speech Consortium test results indicated that at least one channel vocoder was somewhat superior to all LPC's. However, it should be noted that the LPC algorithms did compare quite favorably so that should the eventual cost of practical implementation for LPC prove to be appreciably less than for channel vocoders, it could remain a highly competitive algorithm.

Without exception, vocoders require the careful adjustment of a large number of parameters to optimize their performance. Since such adjustments tend to be made based on a limited number of speakers, they may not be the best adjustments for a more broadly chosen set. Since LPC has not yet evolved fully in this respect, we can anticipate that there will be substantial future improvements.

In Section IV, we discuss directions for future LPC research and the resultant implications for low cost implementation.

At present, a typical LPC system can be subdivided as follows:

1) Analog filter for pre-sampling and possible frequency equalization.

2) Sampler and quantizer at approximately 8 khz and 12 bits.

3) A correlator.

4) Computation of the LPC synthesis parameters.

5) A pre-pitch filter.

6) A pitch detection algorithm.

7) Coding of the pitch and spectral information for transmission.

8) Decoding of this information upon reception of the transmitted bit stream.

9) An excitation generator.

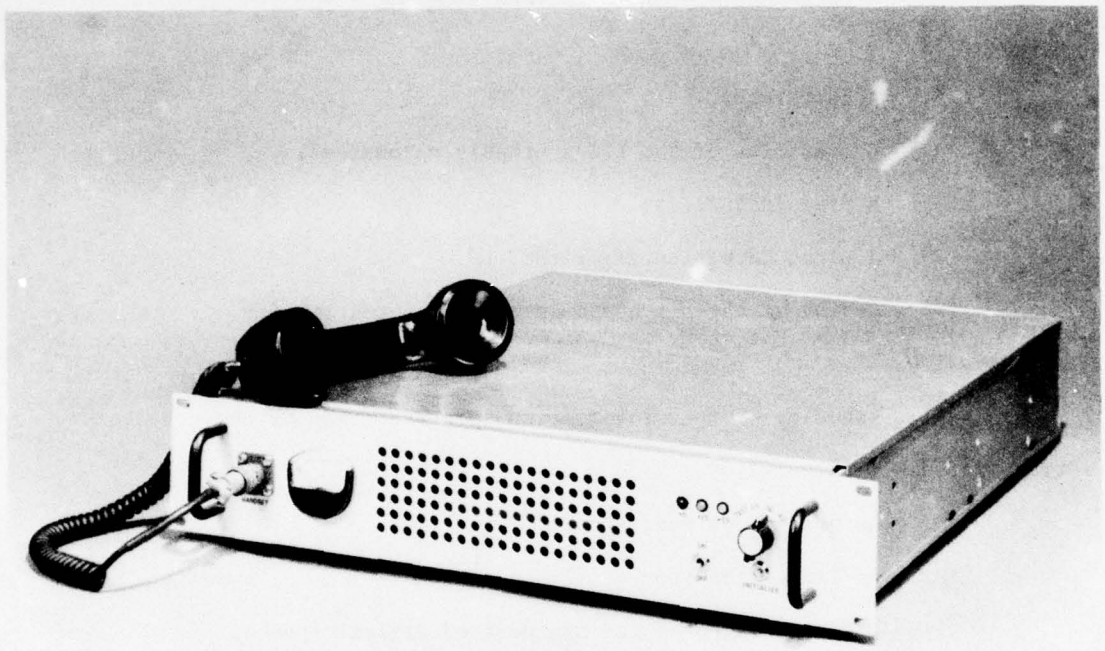
10) An LPC synthesizer.

11) D/A conversion of the synthesized digital speech.

12) Post sampling and de-emphasis filtering.

Detailed implementation of these subdivisions vary but, with the exception of the pitch detection algorithm these variations should not have great impact on the overall size, weight and cost of a complete system.

Several complete hardware realizations of the LPC algorithm already exist; a photo of the LPCM (linear predicitive coding microprocessor) is shown in Figure 1. This device is discussed in detail in the referenced report[2], which also describes what it takes to build a complete full duplex LPC vocoder with technology that is about 2 years old. Also, a good estimate can be made of the cost, size and weight of a new version based on today's technology. A 50% savings in hardware is predictable so that a new LPCM (same algorithm) would be somewhat less than half the size and weight of the device shown in Figure 1.



P246-21

Fig. 1. Linear Predictive Coding Microprocessor (LPCM).

Given a moderate investment in LSI, today's technology offers substantial potential improvement in the size and cost of LPC devices. For example, Litton presented a 20-30 chip LPC microprocessor structure based on 2 LSI chips. In reasonable quantity (over 10000) such a processor should cost well under $1000.

The Litton approach has the advantage of some flexibility due to the embodiment of all the algorithms within a microprocessor framework. One can speculate that (given appropriate LSI development) a more special purpose design could lead to even greater size reduction. The contrast between the two approaches can be seen in Figures 2a and 2b. In Figure 2b, the more time consuming operations have been specialized. It is safe to say that the correlator and synthesizer can each be implemented as single chips. All of the analog filtering can be implemented on a single chip. The resultant microprocessor in Figure 2b could then be strictly a low speed MOS device with an overall chip count of about ten.

B. Channel Vocoder

The channel vocoder has a long history, dating from the mid 1930's. The original invention, at the Bell Laboratories, was viewed by its inventors as a research curiosity, albeit a profound one. From a fundamental point of view, the channel vocoder demonstrated once and for all that speech could be analyzed and then resynthesized in a satisfactory way. Shortly afterwards, the military users adopted the vocoder concept as a means of obtaining secure voice transmission over the switched telephone network. By the 1950's, the role of the channel vocoder as a device for bandwidth reduction was well appreciated. In the 1960's there was renewed research activity leading to improved pitch detection and the notion of spectrum flattening to improve spectral fidelity. As this effort progressed, it became clear that the cost and complexity of such devices was then too great for widespread use. Present day technology appears to have overcome this obstacle. ARPA is presently supporting an effort by Texas Instruments to fabricate specialized chips for spectral analysis and synthesis of speech.

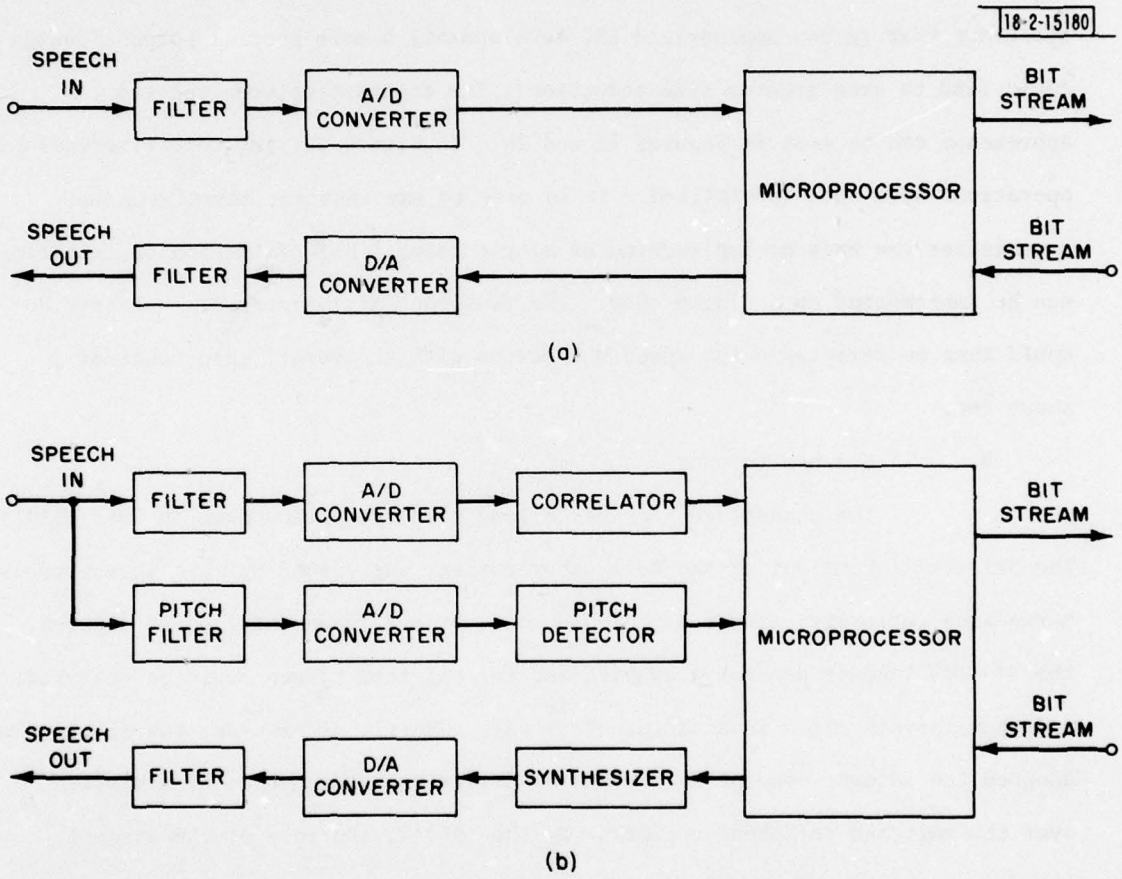


Fig. 2(a&b). Two implementations of an LPC design.

This should lead to a complete channel vocoder with the structure shown in Figure 3, and a form factor of one fifth that of Figure 1. Further straightforward efforts using today's technology can shrink the pitch detector to one or two chips and reduce the entire system electronics to a single 7" x 7" card. Such a device should then be producible at about the same cost as the LPC structure of Figure 2b.

It is interesting to note that a subdivision of a channel vocoder differs from the LPC subdivision listed in Section A only in Items 3 and 10. In Item 3, the correlator is replaced by a spectral analyzer and in Item 10 the LPC synthesizer is replaced by a channel vocoder synthesizer.

C. Homomorphic Vocoder

Since the speech signal can be treated as the convolution of an excitation function and a vocal tract filter function, it follows that an approach to speech analysis is deconvolution. In the 1960's, research workers at Bell Laboratories developed a pitch detection algorithm based on measurement of the cepstrum which, in effect, performs an approximate deconvolution of the speech signal. This was followed by Oppenheim's development of a complete vocoder system based on this principle. In the late 1960's, this system was simulated. More recently a real time simulation has been developed and some early results obtained on intelligibility scores relative to the LPC and channel vocoder systems. These results are quite encouraging and appear to establish the homomorphic analysis and synthesis as a method deserving of further algorithmic research.

The recent real-time effort has been inspired primarily by the advent of CCD technology with its promise of compact high resolution Fourier transform devices. Assuming the existence of such devices in the near future, one can imagine that a homomorphic system can be embodied in the structure of Figure 3, with the analyzer and synthesizer consisting, respectively, of homomorphic analyzer and homomorphic synthesizer.

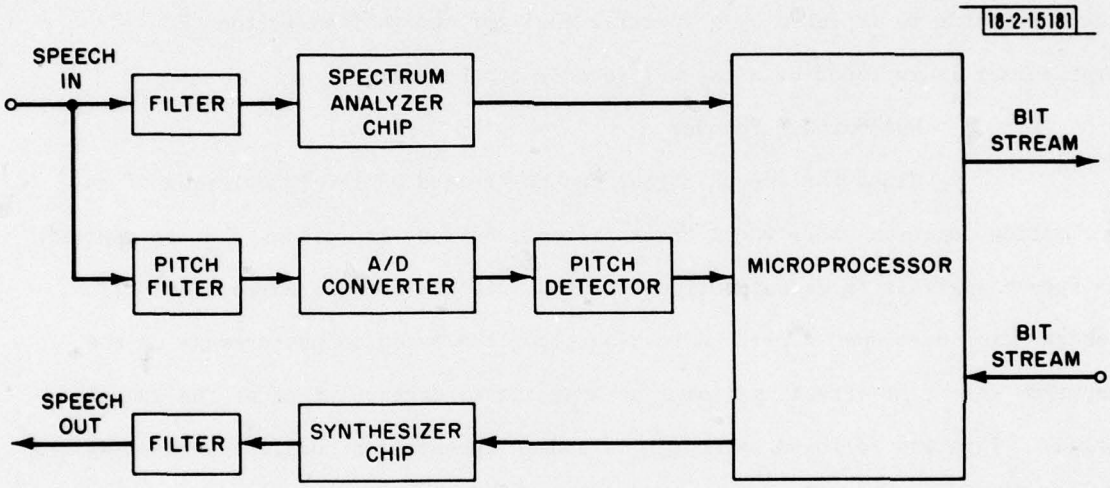


Fig. 3. CCD implementation of a channel vocoder.

IV. ADVANCED VOCODER CONCEPTS

The vocoder systems described in Section III represent viable speech processors for a great variety of today's speech communications situations; yet there is room for substantial improvement and further development. In this section we outline some possibilities along these lines and speculate as to what would be needed in device technology development to turn these possibilities into practical realities. To this end we will discuss five somewhat interrelated areas:

Robustness

Compatibility with advanced networks

Adaptiveness

Ultra-low rates

Improved quality

A. Robustness

We can identify three major robustness problems to which present-day vocoders appear to be particularly vulnerable. The first, perhaps more serious from a military viewpoint, is the vulnerability to background noise. In many practical applications (inside tanks, on ships, in airplanes, in a command post with high level background noise) the quantity of noise encountered is sufficient to significantly lower vocoder intelligibility relative to the performance of wider bandwidth systems such as CVSD and APC. The reason for this can be found in the methods of analysis-synthesis intrinsic to all narrowband vocoders. These methods start with the assumption that the signal to be processed is a vocal cord or vocal noise excitation convolved with the vocal tract impulse response. Clearly, when large amounts of noise are added, or several speakers talk simultaneously or when there is added non-vocal background sounds such as music, the model loses its accuracy. At the present writing, research activity is being directed towards this problem but fundamental experiments need to be carried out before the best possible set of solutions beings to emerge. The first such experiment

should consist of the categorization and assessment (primarily in terms of spectral content) of the various practical noises that are encountered. A next experiment should be the separation of detrimental pitch effects and spectral effects caused by the noise. A third experiment should explore the limits of microphone solutions, such as noise cancelling and throat microphones. Given such data, the application of maximum-likelihood techniques should then lead us to the "best" solutions (in a practical sense) that can be obtained. Such techniques require more processing capability and this leads us to an important general proposition; namely, that improvements in vocoders will almost always entail more computational capability. This fact delineates the need for maintaining structural flexibility in future designs and, concurrently, the need for further advances in device technology so that future terminals do not become prohibitively expensive.

In military situations, the background noise problem is further aggravated by certain tactical and strategic requirements. If, for example, there is a requirement to use spread spectrum techniques to counter jamming or other interference, a low rate speech terminal is needed, but the lower rate increases the background noise vulnerability.

A second robustness issue is the vocoding of telephone speech. Clearly, it will be many years before the additional cost of installing a vocoder in each individual handset is insignificant. By designing the network so that a vocoder services a multiplicity of terminals, economy is attained. This can involve the vocoding of speech that has passed through a carbon button microphone and one or more telephone exchange installations. The phase distortion, change in frequency response, non-linear effects, cross-talk and noise introduced by many telephone channels often deteriorates vocoder quality to an unacceptable level. Present thinking has focussed on the pitch problem as the major problem for telephonic vocoding and to this end there is a high degree of activity towards robust pitch detection. Based on

present simulations, such algorithms appear to require 2 to 4 times as much computational power as the algorithms used in present vocoder implementations. However, the present robust algorithms may still not suffice and a conservative estimate would point towards a required increased capability of an order of magnitude. It has been common knowledge for a long time that a reasonable sophisticated human observation of either the speech wave or fine grain spectrum results in very good pitch detection but as yet little attention has been paid to categorizing the specific algorithms that the human observer brings to bear on the problem.

Channel errors constitute a third major source of vocoder degradation. For example, CVSD at 16 kbps remains reasonably intelligible for channel error rates as high as 10% whereas a vocoder may be hopelessly degraded before the channel error rate reaches 5%. Again, this is an expected result caused by the vocoder's specialized modelling of the synthesized speech. Perception experiments are yet to be performed that relate loss in intelligibility to errors in specific parameters. Given such data, one can imagine methods of selective redundancy encoding of the perceptively crucial parameters. For example, it is probably true that pitch transmission errors should be kept to a minimum. More generally, the channel error problem points at the overall terminal design and the connection between the design of the speech processor, privacy device, and modem. The latter is covered in Section V.

B. Compatibility with Advanced Networks

Within the past decade digital data networks have mushroomed. Two interesting trends pertinent to this report are discernable at this time; one, the trend towards satellite links and two, the trend towards incorporation of voice and data into the same network. A variety of documents deal with these subjects.[1] These strongly indicate the need for more flexible speech terminals. A component of this flexibility is the ability to change the digitized speech

rate in response to either robustness demands or in response to increased network traffic. This leads, first, to a requirement that network and terminal can communicate control information to each other. This subject will not be discussed here but we will discuss the other requirement, namely, that the speech processor itself be designed in such a manner as to gracefully permit each terminal to allow for multiple rate outputs and inputs. Following are some ideas on multiple rate vocoder algorithms and the resultant increase in hardware complexity.

LPC-APC

Both the LPC and the APC algorithms perform vocal tract parameter extraction in the same way; through measuring correlation and then inverting a matrix. Also, both systems require pitch extraction. To switch from LPC to APC requires the generation of an error signal for additional transmission. The order of the LPC analysis that is used in the APC system is generally much smaller than that used for the narrowband case (4 or 6 versus 10 or 12). There is thus some commonality between the two although not a large amount. In any case, we can safely say that a system of this type can be designed to operate at 2.4, 3.6, 4.8, 8 and 16 kbps. Although such a system has not been assessed as to overall complexity, it is our guess that it would turn out to be 2-4 times more complex than either a present day LPC or APC.

PLPC

Another LPC scheme that shows promise of operating successfully at several rates is PLPC. In this scheme, the speech is first divided into 2 equal bands. A multiple rate algorithm could be devised by using a form of adaptive quantization on the residuals of either the higher or the lower band. One then could have the option of using either the quantized error signal or the conventional pitch excitation (or both) for synthesizing in the selected band.

Channel Vocoder

The channel vocoder lends itself quite readily to multiple rate algorithms. The output of each analyzer channel can be sampled and quantized and used as excitation in place of pitch. The selection of excitation in each synthesizer channel filter can be independent, and in this way one can gradually increase bandwidth (and hence robustness) or steadily decrease bandwidth (and hence allow increased network traffic).

Another aspect of network compatibility relates to tandem operation, a situation that will often arise when new terminals are integrated into an existing network or when it is desired to internet two or more networks. Here again, robustness plays the important role; a high bandwidth system will usually tandem better with an arbitrary system than will a low bandwidth system.

C. Adaptiveness

A common occurrence in vocoders is the relatively great degradation suffered by the system when certain conditions are changed. Some of these important conditions are pitch changes, background noise, changes in average spectrum, telephone system changes for telephone speech, changes caused by different microphones. An adaptive vocoder, for example, could perform pitch dependent spectrum analysis and might detect pitch based on a relatively noiseless portion of the spectrum. It might also perform adaptive pre-emphasis filtering and even adaptive pre and post sampling filtering as well as adaptive sampling. Some of these ideas are presently being tried in research laboratories. For example, the window in LPC has been made adaptive to encompass an integer number of fundamental periods. In the homomorphic algorithm, the window size has been made a function of pitch. The maximum likelihood method effectively measures the noise spectrum and attempts to optimize performance under those conditions.

D. Ultra-Low Rates

In some circumstances the transmission channel can only accommodate rates of under 1000 bps. In others, the presence of a jammer might require that error-correction redundancy be employed; this is best accomplished by starting with as low a vocoder bit rate as possible. This leads to the notion of systems that adaptively select the best combinations of vocoder rate and error coding based on measured channel conditions. Over the years, many concepts for reducing the rate below the standard 2400 bps have been proposed. One method, implemented more than a decade ago, was to simply quantize the channel signals in a channel vocoder more grossly; this resulted in a 1200 bps system. Another method, applicable for any of the three basic systems, is to sample the spectrum non-uniformly, so that, for example, during sustained vowels, the resultant slowly varying spectrum would be sampled less frequently.

A fundamental technique for further bandwidth reduction is formant vocoding. It is generally agreed that 3 or at most 4 formant frequencies suffice to adequately model the speech signal. Formant synthesizers can be classified as either parallel or serial. In the parallel case, formant frequencies, formant amplitudes and formant bandwidths need to be estimated; this implies the measurement of 12 spectral parameters or as much as is needed for a good LPC analysis. However, in LPC, the physical or perceptual effect of roughly quantizing a parameter is not well understood whereas in a parallel formant vocoder we know to some extent the effects of, for example, keeping formant bandwidths constant or reducing the number of formants from 4 to 3. It is thus fair to assume that a formant system can lend itself to reduction of rate well below 2400 bps.

Parallel formant synthesis has the advantage that zeros as well as poles are automatically incorporated into the model. Experiments have also

been carried out using serial formant synthesis, which corresponds to an all-pole model. Since LPC, also an all-pole model, is reasonably satisfactory, we can anticipate that a serial formant synthesizer can lead to acceptable low rate speech and at the same time avoid estimation of the formant amplitudes. One such approach has been tried at Lincoln Laboratory, wherein pole positions were estimated by solving for the roots of a polynomial that had been obtained via an LPC analysis; this led to reasonably good results at rates of about 1200 bps.

The major problem encountered for both parallel and serial formant vocoding is the analysis procedure. Both the methods of analysis and their implementations are still strongly in the research stage. We estimate that hardware processing power should be increased by about 5:1 to make ultra low rate vocoding a viable practical field.

Finally, a futuristic vocoder structure might be termed an 'articulatory' vocoder. Here the method of analysis consists of extracting parameters corresponding more precisely to articulator (tongue, lips, velum) motion in an actual vocal tract with the synthesizer presumably being an electrical analog of these parameters. It is too early to predict where this research might lead.

E. Quality Improvement in Present-Day Systems

It is possible to identify three rather straightforward ways of improving vocoder performance and to estimate the increase in hardware capability needed to realize these improvements. The first improvement is in dynamic range. For a single speaker, speaking unemotionally, the sounds of English will vary in volume over a 40 db range. If emotion is included, such as in theatrical speaking, the overall range can easily exceed 60 db. If, in addition,

variations among speakers are included, then an overall range of 70 db is probably a safe figure. Overcoming this problem with digital implementations leads to automatic volume control, 16 bit A/D and D/A converters and, possibly, floating point computational methods. Depending on the precise problem and methods used, we can imagine that hardware capability needs to be increased from 20% to 300%.

A second improvement can be made in the interpolation of the parameters. In most digital implementations, the pitch and spectral parameters are computed every 10-20 milliseconds. Since parameter computation is subjected to statistical fluctuations (depending, for example, on the phasing between the window and the vocal cord excitation), greater accuracy may be attained if the computations are made more often (say, every 3 milliseconds) and then interpolated either linearly or non-linearly. Such a capability, running in real time, requires about a 4:1 hardware speed improvement.

The third improvement is in the vocal tract model. For example, the channel vocoder synthesizer contains fixed poles and controllable zeros whereas the LPC synthesizer contains variable poles and no zeros. We feel that improved naturalness would result from a system wherein both poles and zeros could be varied. For example, a channel vocoder analyzer could be preceded by an LPC predictor, which allows channel vocoding of the error signal. The synthesizer would then consist of an all-pole model followed by a channel vocoder model, thus incorporating both poles and zeros. Other approaches, such as homomorphic LPC, have been proposed as encompassing a pole-zero model. This work requires further research and will probably require a threefold to fourfold increase in processor capability.

V. SYSTEM ASPECTS AND SPEECH TERMINALS

The bulk of future military voice communications will be conducted over wideband digital facilities such as satellite channels, line-of-sight radio links, coaxial cables, etc. Signalling mechanisms will include both packet-switching and circuit-switching techniques, the use of variable-rate and/or embedded coding for traffic flow control, and the widespread use of digital encryption for communications security. Voice connections encompassing more than one networking strategy will be commonplace, with internetting protocols permitting direct voice and data communication between users in a variety of networking situations. The design of speech terminals for large scale use in wideband and internetted environments constitutes a major challenge in system design. In this section we identify some desirable properties of such terminals and we speculate on the nature of future terminal designs.

A fundamental difference between voice and data traffic types is that while the loss of a packet or even a bit is virtually intolerable for data transmissions, the quality and intelligibility of voice communications is practically unaffected by occasional dropouts. A more important requirement for voice is to minimize overall communication delays. These distinctions are generally reflected in the design of network voice protocols, in which the guaranteed delivery of a packet is often relaxed in exchange for reduced network transit time. The actual mechanism through which this tradeoff is effected will vary depending on the specific network structure, but in general one can assume that a common element in future digital voice systems will be that speech traffic will be characterized by occasional missing segments.

Network voice protocols can usually be designed to minimize the perceptual effects of lost data. This generally requires detailed knowledge of vocoder frame structures and synchronization requirements in addition to specific network packet handling strategies. Thus, while voice protocols can mitigate the effects of data dropouts on speech perception, they have to be custom tailored to individual voice algorithms and particular network

characteristics. It is doubtful that a protocol designed for use in one network will perform acceptably when handling traffic that has been generated in some other system and relayed via an internetting protocol. It is also unlikely that a low-cost voice terminal designed for use in a given network, and embodying voice protocols for that network, will be capable of functioning properly in a different networking environment. We are thus faced with the possibility of different terminal designs for use in different networks or with a requirement for very flexible and perhaps unnecessarily expensive terminal configurations.

The above notions lead us to speculate that voice terminals for use in future systems should probably be robust with respect to the occasional loss of input data in a way that is independent of local network protocols. In other words, voice terminals should be able to sustain the loss of arbitrarily long segments of input without requiring prolonged resynchronization intervals. By including this property in a voice terminal, one essentially obviates the need for network voice protocols that are speech algorithm-specific, at least with respect to packet-loss effects. A further separation of speech and networking functions can be effected by performing silence detection and other TASI-like functions in the speech terminal instead of in network hosts. Terminals designed according to this concept may then intercommunicate with each other without specially tailored network protocols, and will therefore be well-suited for use in the wideband internetted systems of the future. In addition, terminal costs should be lowered by virtue of a more uniform design that is compatible with more than one networking strategy.

A critical military aspect of voice terminal design is in the area of communications security. Present techniques for decrypting continuous digital data streams do not easily accommodate data dropouts of the type encountered in packet speech systems. Since voice encryption is most appropriately performed

at the speech terminal (thus relaxing security requirements in network switches, etc.), new or modified encryption techniques will have to be developed in order to cope with the variety of data loss effects that might arise in different networks or in internetted situations. The Black Crypto Red (BCR) concept is an example of one approach to this problem for packet switching systems. In voice systems problems of crypto sync and vocoder frame sync in the face of lost packets might both be handled in a unified way, and in a manner that is independent of network specifics. The notion of selectively encrypting portions of the data stream might be explored in this context.

Present day packet speech activities are restricted to a limited number of users in experimental configurations. Problems of dial-up, call termination, and the interconnection of large numbers of secure digital voice terminals have not yet been fully addressed. Using the standard analog telephone instrument as a reasonable functional model for a voice terminal, we conclude that a terminal should properly contain a dial-up mechanism in addition to a vocoder and crypto device. Since connection protocols will probably vary between networks depending on routing strategies, etc., we expect that these should reside in network nodes. Voice terminals, like telephones, might simply provide a dialing mechanism (e.g., touch-tone pad) via which the user can communicate with a network resource at a local access node. Terminals might also generate audio signalling tones (e.g., busy, ringing, etc.) in response to network messages.

A major systems question emerges from the above considerations; namely, given large numbers of similar secure digital voice terminals, each designed to be robust with respect to data dropouts but without detailed knowledge of network-specific protocols, how may these be appropriately connected to a given network access node? Again, using the conventional telephone system

as a model, we note that regardless of the details of the central switching office or the PBX or any other point of entry into the global network to which they connect, all individual telephone instruments are essentially interchangeable. The observation here is that all telephones share a common interface to their respective network access nodes. Attempts to develop similar properties for secure digital voice terminals result in a rather interesting option, viz, that all terminals that connect to a given access node share a common communications protocol with that node; that this protocol be independent of the characteristics of the network in which the node may be a host; and finally, that the communications protocols between voice terminals and access nodes be identical wherever possible. A major implication of structuring terminals and access nodes in this way is that the terminals may be designed for interchangeable use in more than one secure digital voice system. Resulting terminal hardware costs should be minimized by virtue of a common design and because of the absence of complex network voice protocol or call initiation logic in the individual units. Protocols are imagined to reside in the access nodes, which provide network-specific services for all terminals to which they connect.

Viewed in the above context, a system of voice terminals connected to a digital network through a common access node can be thought of as a local network in its own right. The access node, or speech host, acts as a form of gateway between an external wideband system and the local voice terminal net. The local network might be of the circuit-switched or packet-switched type, but in either case its protocols for terminal-to-terminal or terminal-to-host communications can be anticipated to be especially simple and amenable to realization with inexpensive hardware modules. This follows from the fact that the topology of the local net can be chosen to avoid problems

such as packet order inversion, excessive delay dispersion, etc.; and because the bulk of its traffic load will be of a given type (i.e., voice).

A speculation for voice terminal designs of the future thus consists of the following ingredients:

1) Vocoder algorithms that, in addition to being adaptive with respect to acoustic backgrounds, speaker variations and system bit rate requirements, are robust in the face of missing data segments and perform TASI-like transmission and speech reconstruction functions.

2) Encryption/decryption strategies that, like the vocoder algorithms, are robust with respect to data dropouts of indeterminate duration and can cope with interrupted data streams due to TASI.

3) A touch-tone keyboard for dialing and either lights or audio signals or both for relaying call status (busy, ringing, etc.) to the user.

4) Packetization or other formatting of encrypted speech streams and touch-tone codes in accordance with a simple and uniform protocol for terminal-to-terminal and/or terminal-to-host communications.

5) A modem or interface for connecting the terminal to a local network of similar terminals and a central speech host.

A rough estimate of the amount of hardware needed for such a terminal might be on the order of half again to two or three times that needed for the speech processor portion alone. The ultimate integration of a complete secure digital voice terminal onto a set of chips that can be packaged in a conventional telephone set does not appear to be an unreasonable long-term goal.

We have not addressed the functional requirements, computational capabilities, or architectural features of a speech host in this report.

To a large extent these will depend upon the type of network to which it is connected. We assume however, that economy of scale can be achieved by sharing a common resource among many speech terminals. The benefits of localizing network-specific functions in a central host appear to be substantial, especially with regard to overall system flexibility and simplicity and commonality of speech terminal designs.

VI. The Impact of Hardware Technology on Speech Processing

Current and anticipated developments in device technology should continue to have significant effects on the cost/performance ratio of digital speech terminals. Such developments may be divided into two kinds: those of a general-purpose nature which will be driven by the growth trends of the semiconductor industry and those which are more specialized in scope which are conditioned by speech and other signal processing requirements.

The principal areas of technology to be discussed here are:

1) Digital Memories
2) Switched Capacitor Filters
3) Digital CCD's
4) Analog CCD Processing
5) Microprocessors
6) Fibre-optics
7) Nonvolatile Storage

A. Digital Memories

Probably no area of semiconductor technology has received more attention or research dollars than computer memory. Significant efforts are underway to develop RAM's, ROM's and serial memories with higher densities and more bits per chip.

At present 16K RAM's are being installed in quantity with 64K RAM's and 64K CCD's available on a sample basis. One difficulty in reducing package count is that most memory chips are only 1 to 4 bits wide, requiring multiple chips for 16 bit applications.

It is difficult to imagine any special purpose memory design competing with the development of general purpose devices and their large-volume economies. Eventually, it is hoped, a design will emerge which provides either a wider data buss or sufficient speed to allow serial multiplexing of sub-words.

B. Switched Capacitor Filters

The technique of switched capacitors provides a mechanism for fabricating the equivalent of precision resistors where tolerances and temperature sensitivity are controlled only by the relative areas of two capacitors. It is thus possible

to fabricate high Q filters with simple R-C, operational amplifier combinations using relatively little silicon area. Arbitrarily complicated combinations of poles and zeros may also be realized using this technique.

C. Digital Charge Coupled Logic

Digital charge coupled logic has been proposed as a method of combining the high device density and low power of CCD's with the precision of digital techniques. Devices as complex as a 16-bit FFT butterfly kernel have been proposed but the most complex devices actually fabricated are an 8 x 8 multiplier plus 16-bit adder. This technology can only be used effectively where pipelined operation is feasible such that the inherent low-cost, small area interstage storage of CCD's can be utilized to good advantage. General purpose logic or decision making functions cannot be efficiently implemented and any sub-systems requiring such functions must be implemented with other techniques. It has been argued that for straight shift register-logic operations a digital CCD occupies the same area as an analog CCD which has been designed to achieve the same precision, assuming fixed lithographic tolerances. It is not obvious that DCCL will play a significant role in high speed processors. Considerable improvement over present experimental designs is required before it can compete with standard LSI techniques.

D. Analog CCD Processing

The single most significant development in the semiconductor field with specialized impact on signal processing is the advent of analog charge transfer devices which can provide parallel operations on hundreds of points of a sampled signal waveform traversing through a shift register under clocked pulse control. This leads to the natural implementation of both recursive and non-recursive filters.

While for some radar applications the principal design problems relate to achieving sufficiently high clock rates, the problem for speech applications is in achieving the required dynamic range, given errors produced by photolithographic tolerances, change transfer inefficiencies, [illegible] noise and multiplier non-linearities. The principle activity over the next few years w[illegible]be to demonstrate that sufficient

accuracy can be obtained consistent with the high density structures necessary to achieve the equivalent of 2000-5000 taps per chip. The Belgard vocoder to be built by TI for ARPA should provide insight into questions of tradeoffs between density and accuracy.

E. Microprocessors

Regardless of developments in the area of special-purpose filtering and signal processing for speech terminals, future progress is bound to require additonal capacity for logical operations and decision-making. Increasing refinements will include the introduction of heuristics and the development of pragmatic treatments for the most error-prone portions of the compression process. General-purpose monolithic microprocessors of increasing complexity will provide additional computing power necessary to perform tasks beyond arithmetic and data control functions now handled in machines such as the LPCM. Speaker adaptation, noise cancellation, redundant parameter calculations, detection and branching to special routines will be possible as the increasing capability of microprocessors becomes available at no additional cost.

F. Fibre-Optics

The technology which could have the most significant impact on speech processing and particularly bandwidth compression is considerably removed from the signal processing function. Probably the most dramatic advance which will occur in the next few decades is the use of fibre-optic communications to provide a factor of 1000 or more increase in available bandwidth. The advent of channels capable of handling 10 to 100 megabit/sec should considerably reduce the impetus to reduce speech bandwidth to the range of a few kilobits, except where the use of fibre-optic transmission systems are precluded (e.g., airplanes, ships, land mobile units). A number of installations of a few kilometers in length have been demonstrated and components are readily available for both transmitting and receiving at these rates.

C. Nonvolatile Storage

An interesting combination of technologies which has received only brief attention is the use of nonvolatile storage techniques in combination with CCD's to provide adaptive filtering capabilities. Tap weights can be adjusted by varying the non-volatile stored charge in the MNOS dielectric. Changes in charge storage are effected by writing new values into the MNOS devices when required.

VII. SUMMARY

In this report we have defined three major structures for performing low rate (less than 5 kbps) speech digitization; these are LPC, the channel vocoder and the homormorphic vocoder. It is our opinion that evolving technology will make it feasible to produce low cost implementations of present-day algorithms for each of these structures within 3-5 years. Looking to the future, we see a need for more sophisticated algorithms that include features such as robustness, flexibility with respect to new network designs, adaptiveness to speakers, and the ability to operate below 1 Kbps. How this need might be met has been discussed by identifying a collection of promising technologies. These include switched capacitor filters, Digital CCD's, Analog CCD's, microprocessors, and nonvolatile storage techniques.

REFERENCES

1. B. Gold, "Digital Speech Networks," Proc. IEEE, 65, 1636 (1977).

2. E. Hofstetter, J. Tierney, O. Wheeler, "Microprocessor Realization of a Linear Predictive Vocoder," IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-25, 379(1977), DDC AD-A050860.

APPENDIX
TECHNOLOGY ASSESSMENT

The following pages offer a brief summary of ongoing efforts in various establishments in the fields of I.C. technology and narrow band speech processor research and development.

UNIVERSITY OF CALIFORNIA INFORMATION SCIENCES INSTITUTE (ISI) ACTIVITIES

ISI's primary interest in speech processing is in the systems area. Specifically, they are concerned with the future application of low cost speech processing devices to speech communication on the ARPANET or some future network using packet speech concepts. Thus far, ISI's major contribution in this area has been the development of a Network Voice Protocol (NVP) that has the following properties:

1. It permits a connection to be made.
2. The protocol is independent of the type of speech processor used.

ISI hopes to develop a complete protocol for voice conferencing with the above features. As our device technology and the resultant device develops, it should be useful to keep in touch with the ISI people to help develop the correct system properties of speech terminals.

TRW ACTIVITIES

TRW's effort is directed primarily at the digital CCD area. The purpose of going to digital rather than analog CCD technology was based primarily on the fear that basic accuracy problems exist for analog CCD's. Whether this is of concern for speech processors should become clear after the present Texas Instrument's effort is assessed. It was generally agreed that digital CCD's, being stream devices, were not easily configured into a general purpose structure. Thus, CCD's, whether analog or digital, appear to best match specialized speech

processor architectives, so that competition between them depends on how acceptable the analog CCD vocoder quality proves to be.

An example of the trade-offs for digital CCD's was given via a discussion of a 16 x 16 multiplier being designed. A power consumption of about 150 mW was projected, compared to the present 5 watts for the bipolar multiplier being presently marketed by TRW. Processing time was estimated at 200 nsec which compares favorably with the present multiplier. However, because of the pipelining inherent in the CCD implementation, there is an initial delay of 16 clock pulses, thus making it necessary to embed this device in a special structure.

TRW's major digital CCD speech processing effort at that time was towards a paper study of an LPC processor. They had, on paper, worked out digital CCD implementation of pre-emphasis/de-emphasis filters, autocorrelation Levinson recursion, block scaling, bandpass and lowpass filters and Itakura LPC synthesizer. With these modules it was possible to make some paper estimates for the design of a complete Itakura LPC using specialized chips for the above functions, plus microprocessor chips for functions such as pitch detection and coding. Further projections following LSI development of 3 digital CCD chips (1 for arithmetic and 2 for control) could lead, in their opinion, to a complete Itakura LPC on a 5" x 7" board using 43 16-pin components; these estimates are solely for the speech processor and do not include the other terminal functions.

TEXAS INSTRUMENTS (T.I.) ACTIVITIES

The briefing centered around the effort by T.I. to develop a low chip count CCD implementation of Belgard Vocoder for ARPA.

It is estimated that the analyzer can be developed on one CCD custom chip wherein up to 20 filters, log amplifiers, multiplexing and A/D

converters are included. Also, it is possible that pre-emphasis in the case of the analyzer and de-emphasis in the case of the synthesizer could also be incorporated on one chip. Hence, the analyzer-synthesizer eventual chip count should be two. The lowpass filters incorporated on the chip would be similar to state variable filters developed at the University of California at Berkeley. The log A/D would be MOS. A separate pitch detector could be implemented in CCD using the Chirp-Z Transform. A Gold-Rabiner pitch algorithm, possibly could be implemented in a single chip providing a fundamental design in three custom chips plus peripheral circuits. The overall complement would likely be under 20 device elements.

A 500 stage convolver has been designed in CCD using the CZT approach. A 500 point CZT is in design now for a correlator.

DISCUSSION WITH SPEECH COMMUNICATIONS LABORATORY

Emphasis in this presentation was on improved performance rather than low cost. Thus, it was recommended that efforts be directed toward algorithms that extract pitch at a higher rate, that employ pitch synchronous techniques of spectral analysis, that determine voicing more accurately. It was pointed out that the "robustness" problem is still in an early stage of development and good solutions would probably require substantially greater processing power for real time implementation. In essence, this presentation brought forth the same point of view to be found in the main body of this report; namely, that ongoing and future technology be directed not only towards lower cost of existing algorithm but also towards more practical realizations of the more sophisticated algorithms being and to be developed.

DISCUSSION WITH LITTON

Although Litton informed us that they have withdrawn from the speech processing business, they cordially agreed to give a presentation of their (unsuccessful) Quintrell proposal. This was to develop two basic LSI chips, one for arithmetic and one for control and around this to configure a complete digital LPC using bipolar technology. In this sense, the concept is analogous to that of Lincoln's LPCM work. In the Lincoln effort, the arithmetic and control chips of the AMD 2900 series were used as the building blocks, and this led to a complete LPC device of about 150 chips. The Litton people, by creating their more dedicated LSI chips, hoped to implement the LPC algorithm with 21 chips. They then projected an eventual cost (in lots of 10,000) of $480 per device.

UNIVERSITY OF UTAH ACTIVITIES

The main emphasis here was on advanced algorithmic research, primarily in the area of improved performance in the presence of acoustic background noise.

UNIVERSITY OF CALIFORNIA AT BERKELEY ACTIVITIES

This was, in our opinion, one of the more significant presentations since the emphasis was on innovative analog MOS technology concept that could be directly applicable to speech devices. Whereas CCD's can be used to efficiently implement non-recursive sampled-data filters, these analog MOS devices could implement recursive sampled-data filters. Such an addition should greatly expand the implementation capabilities of present and future speech algorithm.

Further work at UC-B includes an analog 8 x 8 multiplier to be implemented within 10 square mils of surface area, a 10-bit A/D device which is limited by photolithography techniques, an anti-aliasing

filter with charge transfer devices, an analog-to-digital converter which operates at a rate of 10 bits per 10 microseconds, an 8 kHz/8-bit log PCM coder including a compander, a DFT and other implementations using recursive and state variable filters.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY ACTIVITIES

MIT is conducting research on speech algorithms, with emphasis on spectral methods of speech analysis. Motivation for this work was based, first on the belief that more accurate spectral analysis is the correct path towards improved vocoder quality; and second, that foreseeable technological advances would make implementation of these more sophisticated methods economically feasible.

GEORGIA TECH ACTIVITIES

A discussion of methods for performing LPC using analog devices, CCD's and possibly digital attenuators was conducted. Efforts at Georgia Tech center around three areas: a) very low bit rate (up to 100 b/s); b) LPC/CVSD tandeming; and c) objective/subjective quality measures. The bulk of the work was devoted to improving the LPC process.

WESTINGHOUSE ACTIVITIES

Westinghouse has a fairly extensive program in applications of analog CCD technology. In the speech area a fundamental Itakura algorithm implementation is being constructed based on CCD recursive filters, CCD adaptive filters and a programmable CCD filter. The structure has been assembled in breadboard form at this time.

BOLT BERANEK AND NEWMAN (BBN) ACTIVITIES

Several areas were suggested for speech processing improvement. It was recommended that robustness be increased for current algorithms, for situations involving acoustic background noise, telephone environment,

channel errors and tandeming. They believe that algorithms could be tailored to evolving new technology; for instance, CCD's. New algorithms should be developed to improve quality and/or decrease bit rate.

The BBN work involves improved vocoding recognition; helium speech unscrambling; and hearing aids. Also source parameter excitation and more sophisticated methods of spectral analyses are being studies. More work is needed in voiced/unvoiced decisions.

Variable frame rate offers a method for obtaining increased speech quality. For a given average rate, variable frame rate quality is better than obtained with fixed rate. The transmitted rate can be fixed at the average by buffering with no loss in quality. The bits thus saved can be used for error correction and overhead functions such as synchronization.

Another area covered was piece-wise LPC. In this approach, the top-half of the frequency band of speech is folded over and the bottom half of the frequency band can be transmitted by a variety of sampling quantizing techniques or, if the rate needs to be lowered, by LPC analysis.

RCA ACTIVITIES

The primary focus of RCA speech-related device technology has been in silicon-on-sapphire (SOS). They feel that the dimension of their SOS devices could be reduced through Boron diffusion, and that this could lead to a $1000 narrow-band speech processor. Their present SOS microprocessor is called ATMAC, which has already been used to simulate the LPC algorithm. This microprocessor employs two basic chips, a data excitation unit and an instruction and operand fetch unit. They are also developing an arithmetic function unit capable of a 16 x 16 multiply in 700 nanoseconds. The complete LPC operation consumes about 10 watts.

DISCUSSION WITH SYLVANIA

Sylvania took a rather outspoken attitude against custom LSI development. Their argument was based first, on the belief that new methods in speech processing will continue to be developed, and that this implied a dependence on programmable devices. Second, they feel that the development rate of commercial LSI is so great that the future (programmable) speech terminal will be small and cheap.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ESD-TR-78-284

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): A Study of Future Directions for Low Rate Speech Processor Research, Development, and Implementation

5. TYPE OF REPORT & PERIOD COVERED: Technical Note

6. PERFORMING ORG. REPORT NUMBER: Technical Note 1978-43

7. AUTHOR(s): Bernard Gold, Theodore Bially, Jack I. Raffel

8. CONTRACT OR GRANT NUMBER(s): F19628-78-C-0002

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Lincoln Laboratory, M.I.T.
P.O. Box 73
Lexington, MA 02173

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Program Element No. 62706E
Project Code 9P10
ARPA Order 2006

11. CONTROLLING OFFICE NAME AND ADDRESS
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

12. REPORT DATE: 15 November 1978

13. NUMBER OF PAGES: 44

14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*
Electronic Systems Division
Hanscom AFB
Bedford, MA 01731

15. SECURITY CLASS. *(of this report)*: Unclassified

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

narrowband, vocoder, technology, LSI, algorithms, hardware

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

This report summarizes the findings of a study which was undertaken to identify promising areas for future research and development in the narrowband voice terminal area. The objective of the study was to relate device technology developments to voice terminal concepts and designs, in order to better understand the steps that would eventually lead to very small and inexpensive narrowband terminal implementations. The overall conclusion derived from this effort can be summarized as follows: Given the present and near future advances in technology it is predictable that present-day vocoder algorithms will soon be implementable as low cost compact devices. How soon this comes to pass does not depend on any new technological breakthroughs but rather on the direction of investments in LSI development realizable with present device technology. Future systems will require terminal capabilities that exceed those of current designs, and these will result in commensurately more complex hardware implementations.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*