ESD-TR-78-289

(12) LEVEL II

Semiannual Technical Summary

Information Processing Techniques Program

Volume I: Packet Speech Systems Technology

30 September 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 2005).

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

FOR THE COMMANDER

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

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

INFORMATION PROCESSING TECHNIQUES PROGRAM

VOLUME I: PACKET SPEECH SYSTEMS TECHNOLOGY

SEMIANNUAL TECHNICAL SUMMARY REPORT
TO THE
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

1 APRIL – 30 SEPTEMBER 1978

ISSUED 9 FEBRUARY 1979

LEXINGTON MASSACHUSETTS

ABSTRACT

This report describes work performed on the Packet Speech Systems Technology Program sponsored by the Information Processing Techniques Office of the Defense Advanced Research Projects Agency during the period 1 April through 30 September 1978.

CONTENTS

INTRODUCTION AND SUMMARY

Low-cost, robust, flexible narrowband speech terminal development has long been an important DoD activity. Lincoln Laboratory is directing its attention toward the low-cost question by studying and developing algorithms that lend themselves to CCD (charge-coupled device) implementation, and that yield good speech quality and intelligibility for a broad class of speaker types. Two activities directed toward these objectives are reported in this Semiannual Technical Summary. The first is a cooperative effort with industry that will result in a CCD-based channel-vocoder implementation. The second is an algorithm development activity that is aimed at improving the performance of high-resolution Fourier-transform-based algorithms for a wider class of speakers. The algorithms can be implemented using CCD chirp-Z transform devices. A pitch-adaptive algorithm has been developed that is an outgrowth of earlier Lincoln work on a homomorphic vocoder. The algorithm uses a piecewise linear approximation to the log spectral envelope function, and can accommodate a spectral subtraction type of noise-removal algorithm as part of its internal structure.

Variable-rate speech terminals offer the possibility for designing digital networks in which voice traffic can adapt to variations in network loading as a function of time. An effort has been initiated toward developing a multiple-rate speech algorithm that is compatible with an embedded coding or "bit-stripping" adaptation strategy. The algorithm is based on a combination of channel vocoder and subband coder concepts, and yields improved robustness and better voice quality as the bit rate is decreased.

Lincoln is supplying hardware and software to support voice conferencing as part of the ARPA Atlantic Packet Satellite Experiment. Hardware subsystems have been provided to the Norwegian Defense Research Establishment (NDRE), University College, London (UCL), Bolt Beranek and Newman, Inc. (BBN), and COMSAT Laboratories. SATNET conferencing capability was demonstrated in May 1978 in an exercise involving the UCL, NDRE, and BBN sites. Current efforts are directed at the implementation of software in support of internetted conferences with some participants on SATNET and some on ARPANET.

PACKET SPEECH SYSTEMS TECHNOLOGY

I. CCD/BELGARD CHANNEL VOCODER

A. Simulation/Test Facility

A version of the real-time LDSP CCD/Belgard simulation has been prepared to study the behavior of this type of vocoder in the packetized ARPANET environment. The software is arranged to exchange a 7-word buffer (16 bits/word) with the PDP-11/45 host every 20 msec. Coded transmission parameters are packed into the buffer in a format which varies with bit rate, as shown in Fig. 1. This system should afford a flexible mechanism to assess the impact of such issues as silence detection and frame fill-in strategies on vocoder performance. Experimental details are still being formulated.

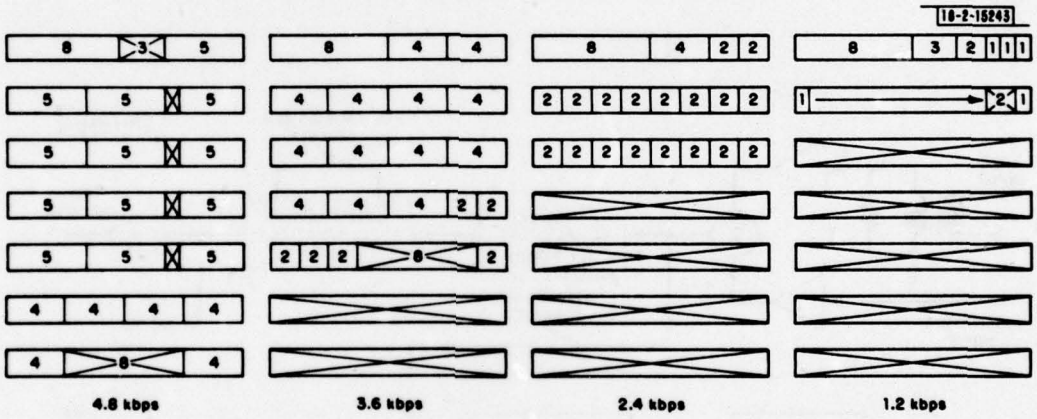


Fig. 1. Packing formats.

Further variants of the real-time simulation are being created to exercise the custom NMOS chips upon their arrival from Texas Instruments. Present plans call for testing of each device independently of the rest of the vocoder during the initial phase. An LDSP computer will be used to simulate the presence of the rest of the system, and will provide an appropriate control/data interface for the device under test. In this way, the actual performance of each chip can be easily compared with expectations in a controlled way.

B. Microcode Status

Four TMS9940 programs have been written, assembled, timed, and debugged via the TI-supplied non-real-time simulation as listed below:

(1) Adaptive peak processing (pitch subsystem),
(2) Final period selection and voicing decision (pitch subsystem),
(3) Encoder/transmit controller, and
(4) Decoder/receive controller.

18-2-14118-1

	TRANSMIT	RECEIVE	PEAK PROCESSING	PERIOD SELECTION
EAROM (bytes)	668 (33%)	730 (36%)	440 (21%)	348 (17%)
RAM (bytes)	119 (93%)	124 (97%)	80 (63%)	38 (30%)
COMPUTATION TIME (W/C-msec)	6.19 (31%)	8.47 (42%)	15.6 (78%)	3.9 (20%)
I/O PINS USED	21 (66%)	15 (47%)	26 (81%)	32 (100%)
INTERRUPTS USED	3	3	2	2

Fig. 2. Microcomputer resource utilization.

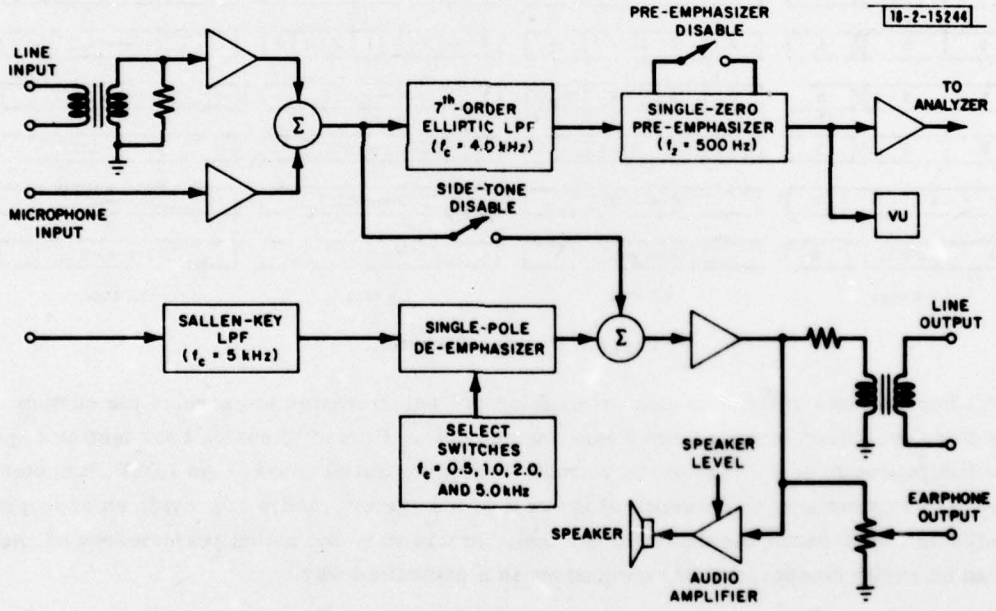


Fig. 3. Audio conditioner.

These programs have been found to operate as intended and are compatible with real-time and memory space constraints. A summary of microcomputer resource utilization is shown in Fig. 2. The only area of concern is the real-time figure associated with the adaptive peak processing. This is a highly waveform-sensitive number and can be expected to vary considerably from speaker-to-speaker. However, exhaustive simulation of the pitch-processing algorithm as reported earlier seems to support the accuracy of this estimate as a worst case.

Some investigations of lower data-rate coding schemes have been conducted parallelling those carried out at JSRU in the United Kingdom. A method was demonstrated at 1200 bps which compared quite favorably with the 2400-bps performance. However, this approach was found to require a bit more RAM space than was available in the TMS9940. A suboptimal 1200-bps scheme was developed which involved a simple augmentation of coding tables stored in ROM where there is ample surplus storage capacity. The approach involves reducing the frame period to 25 msec and transmitting 30 bits/frame using largely 1-bit DPCM. Although somewhat inferior to the JSRU algorithm, the performance was judged to be reasonable and the TMS9940 microcode was modified to incorporate the 1200-bps rate.

The microcode development facility is complete and has been tested insofar as possible without the actual devices. This includes all software necessary for transporting cross-assembler object files from the IBM 370 to our PDP-11/45 host, reformatting the files, and downloading to the LDSP computer. The LDSP acts as the driver/controller mechanism for the EAROM loader interface, and the necessary assembly code has been written for performing this task.

A verification feature has been incorporated into the loading process whereby the TMS9940 being written can be commanded to read back its ROM contents. This provides a mechanism for an automatic cross-check with the image file stored in the two LDSPs. Without this feature, it would be very difficult to establish the correctness of the ROM contents upon completion of the loading process. The TMS9940 programs have been modified to accommodate the new feature.

C. Analog Subsystems

The audio conditioner design has been finalized and a prototype version fabricated, debugged, and evaluated. A conceptual diagram is shown in Fig. 3. On the input side, provisions have been made for both tape- and microphone-level sources. The input is coupled to an active anti-aliasing low-pass filter (LPF) which is generically a 7th-order Cauer design. The essential parameters of the LPF are a transition ratio of 24:1, 0.1-dB passband ripple, and −54-dB stopband attenuation. This particular realization is based upon a compact, commercially available dual bi-quad device. The band-limited source is next coupled to a single-zero pre-emphasizer breaking at 500 Hz. The fully conditioned audio is then connected to a VU meter for level monitoring and the CCD analyzer device.

On the output side, the synthesizer device is AC-coupled to a simple 5-kHz Sallen-Key LPF which acts as a post-sampling filter. A single-pole pair characteristic is sufficient for this purpose given the 40-kHz clocking rate of the synthesizer which affords an effective 20-kHz bandwidth. The post-sampler drives a single-pole de-emphasizer LPF with switch-selectable breakpoints. Options available are 0.5, 1.0, 2.0, and 5.0 kHz. An optional side-tone shunt is added to the de-emphasizer output, and the sum fed to an output buffer. Three outputs are

DIMENSIONS	7.5 × 13.75 × 3.5 in.
VOLUME	360 in.^{3}
WEIGHT	5 lb
POWER SUPPLY CAPABILITY	350 mA at +15 V 300 mA at −15 V 1000 mA at +5 V
PROJECTED DISSIPATION	7.8 W
INTEGRATED CIRCUIT COUNT	VOCODER BOARD − 18 PITCH BOARD − 28 46
COMPONENT BOARDS	TWO 60-SOCKET WIRE WRAP (6.25 × 7.375 in.)
BOARD OCCUPANCY	VOCODER BOARD 46/60 (77%) PITCH BOARD 54/60 (90%)
PANEL FEATURES	FOUR RATE OPTIONS (1200, 2400, 3600, 4800 bps) SELECTABLE DE-EMPHASIS OPTIONAL SIDETONE OPTIONAL LOUDSPEAKER VU METER SYNC INDICATORS

Fig. 4. Physical data.

TABLE I.
ANALOG-FILTER DATA

	Discretes	ICs	Total Slots	Power (mW)
Pre-sample/Pre-emphasis	28	2.25	9.5	420
Pitch Filters (2)	58	3.00	9.0	990
Voicing Filters (2)	21	2.5	6.5	335
			25.0	1745

provided for: fixed-level tape drive, variable-level earphone drive, and variable-level loudspeaker drive. The speaker drive is provided via a compact 0.5-W audio amplifier.

This improved and simplified design has been found to perform satisfactorily and has proven much more compact and power-efficient than originally expected. Therefore, form-factor and dissipation projections have been favorably revised.

D. Mechanical Design

An improved mechanical package design is complete, and fabrication of the prototype is virtually complete. Essentials of the upgraded design are given in Fig. 4. Of particular interest are the improved form factor (372 to 360 $in.^3$), decreased power budget (15 to 7.8 W), and reduced board count (3 to 2, with 17-percent net spare capacity). The accuracy of the power figures is heavily supported by direct dynamic measurements made on the several existing analog subsystems during checkout. This area is the hardest to predict with a high degree of accuracy.

E. Alternative Filter Realizations in NMOS Compatible Technology

The analog subsystems of the CCD/Belgard include five active filters which, along with sundry support components, account for about 25 board sockets and 1.75 W of power (cf. Table I). These filters are: pre-sampling/pre-emphasis cascade, peak processor band passes (2), and the voicing logic high/low passes. It is reasonable to speculate that package count and very likely dissipation figures could be significantly improved if these filters were realized in a more-sophisticated way, preferably in an NMOS-compatible technology. This approach would lend support to the idea of a full VLSI version of the Belgard, including pitch extraction, implementable at a future date as a natural sequel to the present work.

To this end a commercially available, off-the-shelf, 64-stage "bucket-brigade" (BBD) device (Reticon R5602) was carefully evaluated, paying particular attention to the number of extra support components required in a practical configuration. The evaluation circuit is shown in Fig. 5, and additional physical data are supplied in Table II. Evidently, the present state of the art for such devices implies about 23 discretes, 1.5 active integrated circuits, and 547 mW of power (measured dynamically with a 40-kHz clock and a 4-kHz, 1-V p-p CW as the input) required to realize the equivalent of a single active analog design. These numbers compare with 28 discretes, 2.25 integrated circuits, and 420 mW for the pre-sampling/pre-emphasis complex.

Before drawing any firm conclusions, several points should be considered:

(1) It must be shown that a satisfactory FIR design can be developed for each of the five cases of interest given a 64-stage constraint on the filter length. It may happen that differing sampling rates are necessary for each of the five designs to ease the problems with the length limitation. Figures 6, 7, and 8 indicate that reasonable designs can be found for the pre-sampling- and pitch-filter complexes, but that clock rates of 2.5, 5, and 40 kHz are necessary.

(2) If several filters are to be realized, some of the support circuits can be shared. This includes the resistive bias ladder, the clock driver 10-V reference, and (if a common clock is possible) the clock driver itself.

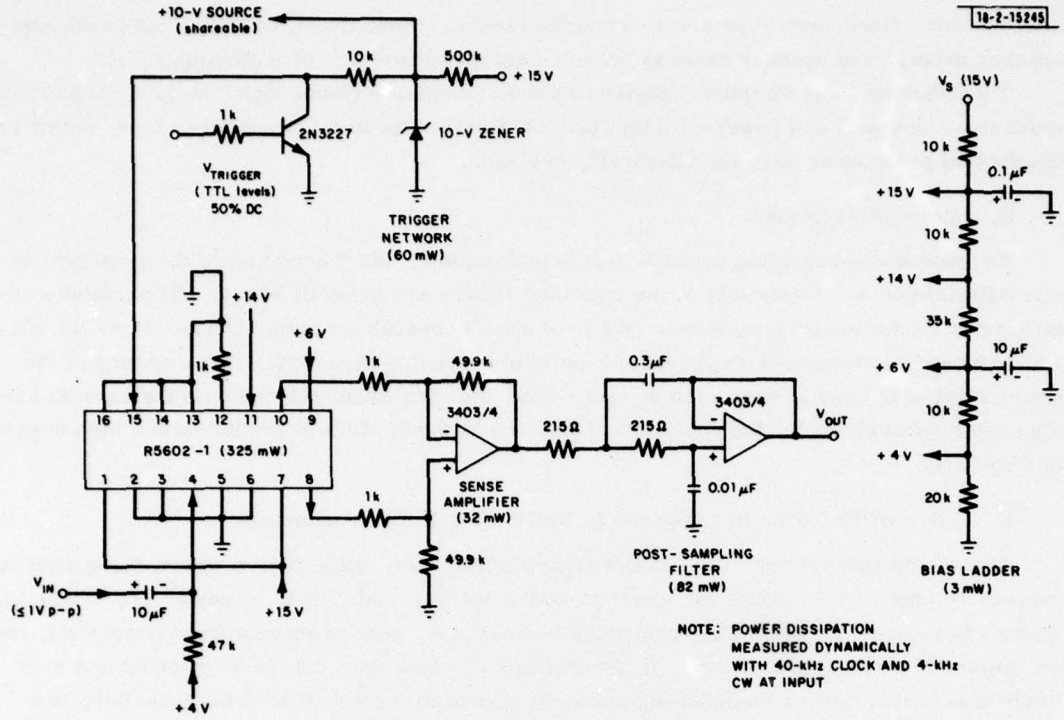


Fig. 5. Reticon BBD low-pass-filter evaluation configuration.

TABLE II
R5602 FIR-FILTER PHYSICAL DATA

	Discretes	ICs	Power (mW)
Low-Pass (Sallen-Key)	4	0.25	82
Bias Ladder	7	–	3
Trigger Circuit	5	–	55
Sense Amplifier	4	0.25	82
R5602-1 and Support	3	1.00	325
	23	1.50	547

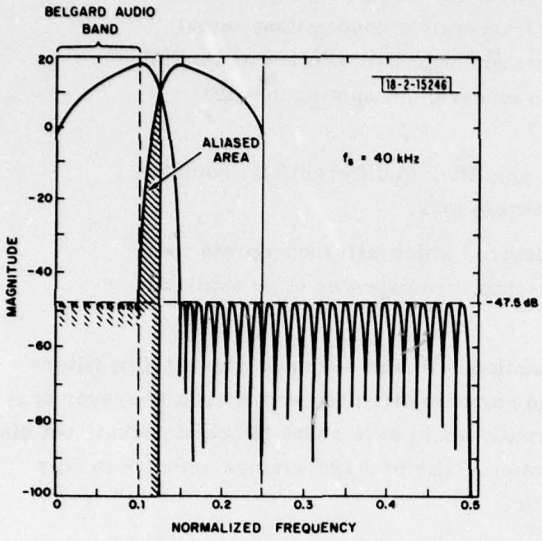


Fig. 6. Pre-sampling/pre-emphasis filter.

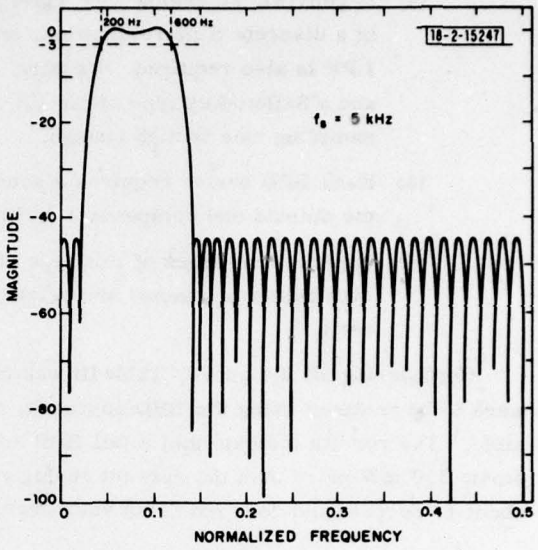
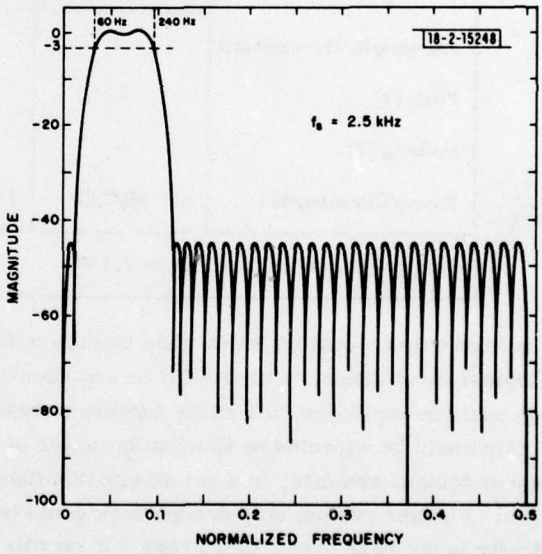


Fig. 7. High-band pitch filter.

Fig. 8. Low-band pitch filter.



(3) In general, an analog LPF must precede the BBD device given that it is a discrete time realization, and frequently a concomitant output LPF is also required. Usually, a single-pole pair section will suffice and a Sallen-Key type of circuit can be used, presuming the BBD sampling rate is high enough.

(4) Each BBD device requires a sense amplifier to differentially combine the outputs and compensate for insertion loss.

(5) Improved versions of this type of device, which will incorporate the bias ladder and sense amplifier on-chip, are expected to be available soon.

Considering all the above, Table III was compiled. It was assumed that all five filters were to be realized using the BBD approach, and common circuitry was shared wherever possible. The results indicate that a full BBD approach would save about 13 board sockets but dissipate 350 mW more than the current analog version. The package savings amounts to only about 13 percent and does not seem very dramatic.

TABLE III
PROJECTIONS BASED ON FUTURE DEVICE

	10-V Reference	Trigger Circuits	LPFs	R5602
Pre-sample/Pre-emphasis	1	1	1	1
Pitch (2)	–	1	3	2
Voicing (2)	–	1	1	2
Power/Discretes/ICs	50/2/0	15/12/0	410/20/1.25	1626/10/5
Total: 44 discretes, 6.25 ICs at 2.1 W				

It is known that a custom NMOS chip incorporating all five filters is well within present technological capabilities. Either CCD or switched-capacitor techniques, or a combination of the two, could be employed, the exact mixture to be determined by appropriate trade-off studies. Such a chip could be expected to save on the order of 25 board sockets and 1.5 W of power over the present design, resulting in a net dissipation figure near 5 W and a 25-percent board area reduction. Further custom chip design work could be done in the pitch computer subsystem, specifically in the peak processing areas. It may be possible to eliminate two 9940's, the pair of sample/hold A/D converters, and sundry control SSIs. The entire pitch computer could well reduce to a half-dozen chips or so.

II. MULTIPLE-RATE VOCODER

The previous Semiannual Technical Summary* reported on a particular scheme of multiple-rate speech processing in which the bandpass-filter outputs of the vocoder analyzer were sampled and quantized to 1 bit. This information could then be transmitted along with the usual channel vocoder information. If these extra bits were stripped by the network via an embedded coding strategy, the receiving speech processor would assume the form of a channel vocoder. The information that reached its destination could be used as a more-robust excitation signal. As reported, this system did indeed improve robustness as the rate increased. Furthermore, this system had the advantage of great simplicity, since no extra filtering was required; and, indeed, the only additional complexity consisted of integer-band sampling and a 1-bit quantization plus a small amount of control.

However, there were disadvantages. The most striking disadvantage was the failure of the system to noticeably improve quality (as the rate increased) for good input speech. Another failing was the inability of the system to do much better than a channel vocoder for severe input spectral distortions, the system improvement in this case being almost wholly due to improvement in the excitation signal. Finally, additional complexity was introduced because spectral flattening was a necessary adjunct of the system architecture.

For these reasons we decided to experiment with an alternate scheme, but still making use of the subband-channel-vocoder hybrid concept. For this experiment we (at least temporarily) accepted the notion of having both a complete channel vocoder and a complete subband coder operating side-by-side. Several channel vocoders had already been implemented; we chose to use the spectrally flattened version that we had already used in the previous experiment. We then implemented a subband coder using frequency-sampling filters as the basic building blocks. By spacing the poles of these filters 25 Hz apart, we were able to design a bank of six bandpass filters with nearly linear phase but with sharp transition bands and low sidelobes that made them suitable for integer-band sampling. Although the pass bands of these filters did not exactly match an integral number of vocoder filters, the match was sufficient to prevent significant spectral distortion when the outputs of the two systems were added.

The switching strategy could now be made quite simple. Whenever it received a particular subband, the receiver would delete the vocoder filter outputs corresponding to that subband by setting the appropriate modulator signals to zero.

This new system has just recently been implemented, so it is not possible presently to give many details of the results. Preliminary runs clearly demonstrate that for good input speech, the output-speech quality improves monotonically with data rate. A few early runs with speech tainted by airborne-command-post noise indicate that robustness for that case is also enhanced.

It is worth noting that this system at rates in the range 7 to 10 kbps looks very much like a voice-excited vocoder and, as such, could be expected to be competitive to other systems in that rate range, such as APC or VELP. Informal listening verifies that this is indeed the case. Furthermore, since APC is particularly vulnerable to channel errors and there is little a priori reason to expect comparable vulnerability for our hybrid system, we can at least speculate that the latter is a preferred system in certain situations.

---

*Information Processing Techniques Program Semiannual Technical Summary, Volume I: Packet Speech/Acoustic Convolvers, Lincoln Laboratory, M.I.T. (31 March 1978), pp. 12-14, DDC AD-B028561-L.

More-comprehensive listening under a variety of conditions is a prerequisite for authorative remarks on system performance. In addition, however, there are significant implementation issues that need to be addressed. The frequency-sampling filters mentioned previously could be used to implement a complete channel vocoder so that the necessity for two separate devices would disappear. However, technology is presently not at the point where such an implementation could compete economically with the CCD approach. Thus, an effort is presently under way to implement a subband system with nonrecursive methods, since such methods lend themselves to CCD technology.

Finally, the new scheme described here should work well with a standard channel-vocoder design that does not include spectral flattening. Experiments ascertaining the correctness of this statement will be run.

III. SPECTRAL ANALYSIS VOCODERS

A. Speaker-Adaptive Vocoder

Informal listening tests had indicated that the spectral envelope estimation vocoder with a Gold-Rabiner pitch detector sounded quite natural at bit rates of 2.4 kbps and above. DRT tests, performed by RADC/ET, yielded the following intelligibility scores:

Vocoding Rate (kbps)	DRT Scores
Uncoded	89
8.8	88
4.8	87
3.6	88
2.4	88

A 2.0-kbps coding which causes only minor degradation to the vocoder has been developed but has received no formal testing.

Speculation that the spectral envelope estimator might be robust with respect to additive noise has yielded two additional developments – noise suppression and an integrated maximum-likelihood pitch detector. Both these processes require an estimate of the background noise. As human speech has silences due to gaps between phonemes and pauses, frequent moments exist where only the background noise is present. Therefore, a time-decaying "minimum" tracker on the spectral envelope will estimate the noise spectrum.

$$|\hat{N}(f)| = \min \begin{Bmatrix} |\bar{Y}(f)| \\ (1+\epsilon)|\hat{N}(f)| \end{Bmatrix}$$

done on a frame-by-frame basis

where

$|\hat{N}(f)|$ is the noise spectral estimate,

$|\bar{Y}(f)|$ is the speech-plus-noise spectral envelope, and

$0 < \epsilon \ll 1$.

The time decay (controlled by ϵ) is required to allow initial adaptation and adaptation to changing noise environments.

If the speech signal and the background noise are uncorrelated, their power density spectra add. Thus, an estimate of the speech spectral envelope can be achieved by:

$$|\hat{H}(f)| = \max \begin{cases} \sqrt{|\bar{Y}(f)|^2 - \alpha |\hat{N}(f)|^2} \\ \beta |\bar{Y}(f)| \end{cases}$$

where

$|\hat{H}(f)|$ is the speech spectral envelope estimate,

α controls the degree of noise suppression, and

$0 < \beta << 1$ limits the musicality of background noise.

This process* greatly suppresses the background noise but places a musical character on the remaining noise. The parameters α and β can be varied to control the amount of noise suppression, musicality of the background, and distortion of the speech.

A maximum-likelihood pitch detector which shares initial signal processing with the spectral analyzer has been developed to allow pitch detection in noisy environments. The likelihood function is computed as follows:

$$\ell(T) = \text{DFT}\left[\frac{w(f)|Y(f)|^2}{|\bar{Y}(f)|^2}\right]$$

for a time-waveform Hamming window 3.5 pitch periods long, where $\ell(T)$ is the likelihood function and

$$0 \leqslant w(f) \leqslant 1$$

$$w(f) = \begin{cases} 1 \text{ for regions of high S/N} \\ 0 \text{ for regions of low S/N} \end{cases} .$$

The pitch estimate $\hat{T}$ is chosen to maximize $\ell(T)$ over the interval 2.5 msec $\leqslant T \leqslant$ 20 msec. The voicing decision is based on the value of $\ell(\hat{T})$ and on the continuity of the pitch estimates. In the clear, this pitch detector performs reasonably well, although not as well as the Gold-Rabiner pitch detector. In noisy environments, however, its performance is superior to the Gold-Rabiner algorithm. Informal listening tests indicate that the algorithm performs reasonably well in both airborne-command-post and helicopter environments.

Work is continuing in several areas. The vocoder imparts a buzzy quality to the speech, and a number of tests are being performed to explore techniques for removing this effect. Alternate synthesizer structures are being investigated. A filter-bank synthesizer (42 channels, 90-Hz spacing, 16-Hz bandwidth simple resonators, antiphase summed) with an echo-removal filter is one of the structures under consideration. Techniques for improving the performance of the maximum-likelihood pitch detector in the clear are being studied. Ultra-low bit-rate codings without undue degradation may be possible, and should also be investigated. Diagnostic Rhyme Test tapes have been submitted to Dynastat for evaluation of the vocoder's intelligibility in several noise environments.

B. Harmonic Pitch Detector

The harmonic pitch detector algorithm has demonstrated potential for providing robust pitch estimation in degraded speech environments, specifically that of the telephone. Efforts of the

*Similar to a noise-reduction process proposed by R. Schwartz and M. Berouti of BBN.

previous six months have centered on a detailed analysis and evaluation of the system's performance in clear speech in the hope of improving several critical weaknesses and identifying any others.

As originally proposed,* the harmonic pitch detector attempts to extract pitch from the distances separating the harmonics present in voiced-speech spectra. To obtain accurate pitch estimates, a scheme was devised in which an FFT was performed on a filtered and down-sampled signal containing exactly 1/9 of the original spectral content (230 to 1148 Hz at a 121-μsec sampling interval). Computational savings were achieved by dividing the overall operation into two filter and down-sample substages employing FIR filters.† Although this multi-stage scheme is computationally efficient, it also leads to an undesirable increase in the effective impulse-response length of the overall filter, since the actual order of the second-stage filter must be multiplied by the preceding down-sampling factor to determine the length of its impulse response at the original sampling rate. The overall impulse-response length of the two-stage design is 19.36 msec, and this size was deemed inappropriate for spectral pitch detection due to its proclivity for incorporating rather large portions of the past speech into the current analysis frame. To reduce the resulting spectral distortions, a single-stage filter was designed whose impulse-response length is 8.7 msec. This modification led to better pitch estimation, especially in those voice segments of speech such as diphthongs, where the spectral characteristics of the signal undergo rapid transition.

A related issue of fundamental importance to the operation of any spectrally based pitch detector concerns the length of the frame to be analyzed. The larger the frame size, the more successful the algorithm is likely to be in extracting low-frequency estimates of pitch. However, during regions of spectral transition where voicing is present, a longer spectral window is less likely to provide useful harmonic information than is a shorter one. For this reason, a 30.5-msec analysis window has been substituted for the original 38-msec window, and informal listening has indicated a distinct improvement in speech quality. However, the new system is not expected to perform as well for speakers with very low pitch.

Another weakness in the algorithm has been observed in relation to speech sounds characterized by low first formants, such as /i/ (beet) and /u/ (boot). The spectrum between 230 and 1148 Hz may contain only one strong harmonic, usually of low frequency. Since the pitch-estimation routine measures distances between pairs of adjacent peaks, erroneous estimates may occur. Although this difficulty can be virtually eliminated by lowering the threshold of acceptance for spectral peaks, such an approach will not be effective for either female speakers or speech corrupted by noise.

Probably the most crucial shortcoming of the original pitch-detector algorithm is the technique employed for discriminating between voiced and unvoiced frames of speech. The method, which relies on pitch track continuity and silence detection, has not demonstrated satisfactory performance in listening tests with clear speech. To augment these measures, a method was devised which searches for harmonic structure in the spectrum to determine whether the speech

---

*S. Seneff, "Real Time Harmonic Pitch Detector," Technical Note 1977-5, Lincoln Laboratory, M.I.T. (26 January 1977), DDC AD-A038542/7.

† R. E. Crochiere and L. R. Rabiner, "Optimum FIR Digital Filter Implementations for Decimation, Interpolation, and Narrow-Band Filtering," IEEE Trans. Acoust., Speech, and Signal Processing ASSP-23, 444-456 (1975).

segment is voiced or unvoiced. The algorithm considers the pitch estimate in conjunction with the magnitude spectrum, and examines the spectral content a pitch distance and half a pitch distance from any peak identified as a harmonic. For voiced segments, spectral magnitudes a pitch distance away from a harmonic lie on a peak, while those half a pitch distance away lie in a valley. The sum of these peak-to-valley distances is expected to be relatively large during voiced segments, and relatively small or even negative during unvoiced segments. This comb-like method has been incorporated in the harmonic pitch detector and has resulted in a noticeable improvement in speech quality for male speakers. However, this measure has not performed reliably in regions of voiced transition where the spectrum, having been computed over a relatively long time interval, no longer contains distinct harmonics and appears very much like unvoiced spectra. Here again, long time windows and excessive filter lengths tend to reduce the usefulness of this measure.

A version of the pitch detector containing the modifications mentioned above has been embedded into a non-real-time vocoder structure such that its performance may be compared directly with that of an identical vocoder using the Gold-Rabiner pitch detector. Informal listening tests have shown that the modified algorithm performs nearly as well as the Gold-Rabiner algorithm for male speakers. Although all tests are currently being performed with clear speech, modifications are being considered only if they may be expected to provide good reliability in a telephone environment. Efforts are presently being concentrated exclusively on the development of a robust-voiced/unvoiced algorithm for the harmonic pitch detector.

IV. SATELLITE CONFERENCING

Lincoln Laboratory's role in the ARPA Atlantic Packet Satellite Experiment is to provide hardware and software to support voice conferencing experiments in the Packet Satellite Network (SATNET). With the delivery of the fourth set of hardware to COMSAT Laboratories in August, all planned hardware is in place and ready for experimental use. The other three sets are located at the Norwegian Defense Research Establishment (NDRE) in Kjeller, Norway, at the University College (UCL) in London, England, and at Bolt Beranek and Newman, Inc. (BBN) in Cambridge, Massachusetts. The voice conferencing experiments are intended to make use of the broadcast stream capability of SATNET which has not yet reached operational status and is now expected to become available in late October 1978. Meanwhile, in order to exercise the equipment and gain experience with SATNET and the ELF operating system in use in the SATNET gateway computers, we have developed software to support both point-to-point speech communications and voice conferencing using existing SATNET capabilities (fixed TDMA). A few experiments and one demonstration have been carried out, but SATNET has been generally unavailable or unusable for speech communication due to development work on SATNET itself. We expect that opportunities for further experiments and demonstrations will occur as SATNET reaches operational status in the first quarter of FY 79.

A. Conferencing Software

The program to support SATNET point-to-point speech communication was extended to handle conferencing using a modification to the SIMP (Satellite Interface Message Processor) software which provided broadcast delivery of packets. This interim program uses the same basic conferencing strategy planned for the broadcast stream capability, but the details of communication between the program and the SIMP will be quite different in the SIMP version which

realizes that capability, and substantial revision of our program will be required when the new version is made available.

Work has started on the implementation of software to support internetted conferences with some participants on SATNET and some on ARPANET. The existing datagram gateways between the networks are not capable of handling the demands of conferencing which require the gateway to have knowledge about the existence and makeup of the conference in both nets. We have designed a specialized conferencing gateway program which will appear as a conference participant in both networks. In SATNET, it will be one of several Conference Control Programs sharing distributed control of a SATNET conference. In ARPANET, it will be a Chairman Program exercising centralized control of an ARPANET conference. SATNET conferees will appear to the ARPANET conference as different extensions at the gateway site on the ARPANET, and vice versa. The gateway program will deal with the special timing and addressing requirements of the SATNET broadcast stream, and will also handle the replication of messages required in ARPANET.

B. Demonstrations and Experiments

The interim SATNET conferencing capability was demonstrated at the Packet Satellite Working Group Meeting in London in May 1978. The conference involved three sites: UCL, NDRE, and BBN. There were some problems with loss of packets in SATNET and a couple of instances of loss of vocoder synchronization, but the demonstration was generally successful. The distributed control program demonstrated its ability to recover control of the conference following the loss of communication with one participant, who subsequently automatically re-entered the conference.

We participated in several experiments undertaken by NDRE to measure the packet delay characteristics of SATNET and internetted point-to-point speech. Support of these experiments required some extension of our speech software to provide additional operating modes to facilitate the collection of measurement data. Round-trip speech delays between NDRE on SATNET and Lincoln Laboratory on ARPANET were observed to be nearly 5 sec.

C. ARPANET Speech

At the request of ARPA, a tape of an ARPANET voice conference was made involving participants at the Information Sciences Institute (ISI), Culler-Harrison, Inc. (CHI), and Lincoln Laboratory. The tape was made using the LPC II variable-rate speech-encoding algorithm.

In order to be able to carry out internetted speech experiments between SATNET and ARPANET, it is necessary to extend the ARPANET voice protocol to deal with the 2400-bps LPC encoding scheme used in the LPCM hardware provided for use in SATNET. ISI has agreed to implement the LPCM vocoder algorithm in their Floating Point Systems processor in order to be able to participate in internetted conferences. This implementation is now reported to be ready for testing.

UNCLASSIFIED

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

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: ESD-TR-78-289-Vol-I	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)*: Information Processing Techniques Program Volume I: Packet Speech Systems Technology	5. TYPE OF REPORT & PERIOD COVERED: Semiannual Technical Summary 1 April - 30 September 1978
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Bernard Gold and Ernest Stern	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: ARPA Order-2606, Program Element No. 62706E, Project No. 8P10
11. CONTROLLING OFFICE NAME AND ADDRESS: Defense Advanced Research Projects Agency, 1400 Wilson Boulevard, Arlington, VA 22209	12. REPORT DATE: 30 September 1978
	13. NUMBER OF PAGES: 22
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

Supplement to ESD-TR-78-366 (Vol. II) and ESD-TR-78-396 (Vol. III)

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

packet speech	coherent integrator	homomorphic vocoding
network speech	voice conferencing	linear predictive coding
syllable scoring	ARPANET	cepstral pitch extractor
convolvers	SATNET	

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

This report describes work performed on the Packet Speech Systems Technology Program sponsored by the Information Processing Techniques Office of the Defense Advanced Research Projects Agency during the period 1 April through 30 September 1978.

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

UNCLASSIFIED

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

207650