Comments on HF Digital Communications
Part 1 -- Link Level Issues

Tom Clark, W3IWI
6388 Guilford Road
Clarksville, MD 2 1029

1. Introduction: This paper and its companion Part 2. wil1 discuss some of the issues involved in improving the throughput and reliability of amateur HF digital communications links. The intent is to present some strategies which should allow for a significant improvement with modest hardware and software investments.

As a member of the ARRL's "SKIPNET" HF "Special Temporary Authority" (STA), I feel very frustrated with the current situation. Even though the number of packet stations around the world have increased geometrically over the past 5 years, the ability of the long-haul HF networks to support the user-generated message traffic has been, at best, a "level" resource.

It is my opinion that there are two fundamental problems:

1. At the link level all digital modes are using rudimentary technology which grew out of standards which were merely convenient and which are far from optimum.

2. At the protocol level, AX.25 (as it is currently used) is incredibly inefficient and needs to be replaced.

I believe that developments in these two areas should proceed in parallel and will treat the topics separately.

2. The HF Radio "WIRE": It has often been stated that radio links are the worst possible "wires" for the carrying of data. This statement is especially true at HF. Modem designers using real "wires" are able to make three simplifying assumptions?

- The signal-to-noise ratio (SNR) is high and signals are stable for long periods.
- Any additive noise present in the system has a Gaussian probability distribution.
- Even though non-line&ties may be present in the "wires", their effects are constant for long periods of time and may be handled by simple adaptive equalization.

Judged by these criteria, even our best VHF/UHF paths have inferior "wires", but at HF none of these criteria are applicable. And yet we attempt to use the wireline Bell 103 modem standards (200 Hz shift, 300 baud, one bit/baud) for HF packet and similar standards (170 Hz shift but with lower baud rates) for RTTY and AMTOR.

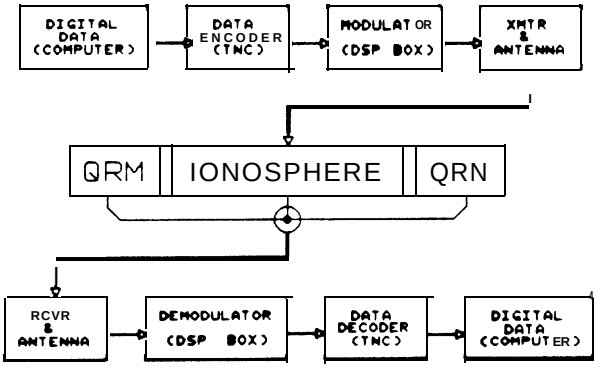
3. Modem Standards -- Old and New: The reason for using the 103-like standards is purely historical. In the early days 103 modems were available (and WB4APR picked up a large quantity on surplus). The 103 modem standards were easy to implement in the Exar XR2211 PLL demodulator in TAPR TNC-1 and TNC-2 and their clones. Manufacturers like AEA (with the PM-1 and later the PK-232) were able to use filter-based demodulators based on their RTTY "TU" products. The AMD AM7910 and TI TMS3 105 single-chip "wireline" modems

provided inexpensive implementation paths for other manufacturers. But easy availability does not make the use of the simple modem technology optimum!

We will soon have Digital Signal Processing (DSP) hardware available. N4HY, W3IWI and others have been experimenting with TI TMS320-10 and -15 hardware for several years; this experimentation has led to several "products" which will be available in late 1990. The initial "offerings" will use the Motorola 56001 (AEA, DRSI and AMRAD) or TI 320-25 (TAPR/AMSAT) DSP chips. The AEA and DRSI hardware will be sold commercially, while AMRAD, TAPR and AMSAT are amateur development groups.

All these "products" are supposed to have well-defined "open architecture" so that many people can experiment with innovative modem ideas. When a new idea is available, the code will be distributed by telephone BBSes, Internet FTP file servers or over the air. I anticipate the early period of experimentation will involve membership in the "Modem of the Week Club"! The important thing to remember is that with either the 56001 or 320-25 we will own "engines" with computing power equivalent to a significant fraction of a Cray Y-MP (when doing well-defined specific tasks) and that our "Personal Cray" will have a cost less than the HF transceiver it is connected to.

4. The Data Path: In order to design an "optimum" modulation/demodulation system (if such even exists) we must consider the overall data transfer system, as outlined in this diagram:



This drawing is intentionally simplistic. The TNC-like data encoding/decoding functions may take place in the host computer or they may be in a separate box (for example, the TNC functions might be in the DSP box, or they may be in the host computer, or the functions may be split). Some data manipulation operations such as forward error correction (FEC) or convolutional encoding may best be handled by the DSP "engine". But for simplicity we shall consider the modulation/demodulation (MODEM) functions separate from coding functions.

In order to try to optimize the modem operations, it is necessary to consider the entire signal path. The cascaded effects of the radios, the antennas and the ionosphere can be considered to be a filter. An optimum demodulator needs to employ the conjugate filter to achieve optimum performance.

The radios and antennas are perhaps the simplest to deal with since their performance is nearly constant with time. James Miller, G3RUH has demonstrated that the combined inadequacies of FM transmitters and receivers can be improved by pre-distorting the transmitted waveform of 9600 BPS FSK signals. In principle, pre-distortion could be used at HF also.

Of more concern are the effects of ionosphere on the signals. Unfortunately radio signals propagated through the ionosphere exhibit considerable time variability and variability with

signal frequency. We shall return to this topic in subsequent sections.

The final area where the HF modem designer's life is made much more difficult arises because of the nature of the additive noise environment. While the wireline modems can assume a nearly Gaussian noise background, HF is plagued with impulsive noise (QRN) from thunderstorms, automobile ignitions, motor starting relays, and even the XYL's blender and garbage disposal. RF interference, both inadvertent and intentional (QRM) is a normal occurrence. The major defense against QRM and QRN is to insure that data protocols are tolerant of dropouts. It is also necessary to have receivers and demodulators which have good dynamic range and which recover rapidly from "glitches". Barry, VE3JF has also argued for the use of frequency diversity to help mitigate against such problems.

5 Bits vs. Bauds: In wireline applications we often see modulation schemes which jam upwards of 20,000 bits/sec through a 2 kHz wide telephone line. Many people *incorrectly* refer to V.32 modems as "9600 baud full-duplex" or TeleBit Trailblazers as handling up to 18 kbaud. It is often said "If such signals can go through a phone line, why don't we use them on the radio?"

The V.32 modems use a complex trellis coding scheme where an in-phase and quadrature phase channel are each amplitude modulated after the bits are convolutional encoded. The resulting complex waveform must handled very carefully if information is not to be lost. Typical $600 commercial V.32 modems use custom DSP chips with 18-bit A/D and D/A converters. The phone lines are equalized by a "training sequence" at the start of each transmission and assumed to be stable thereafter. Such are not the characteristics of our radio links!

The confusion between bits and bauds must be clearly understood. A bit is a piece of data, A baud represents an interval in time to accomplish the signaling.

As a simple example, consider the touch-tone telephone (DTMF) system. Each signaling element consists of one tone chosen from four low tones, and one of four high tones. The two one-of-four signals encode 16 possible states, equivalent to 4 bits. The same scheme could be expanded to having any number (l-8) of tones present, in which case there would be 255 possible states -- almost 8 bits worth of data (in this example, the zero state with no tones present is undefined).

For those who have studied Fourier transforms, a fundamental theorem of spectral analysis is that a frequency can be measured to an accuracy A f in a time interval Δt related by the uncertainty principle

$$\Delta f \, \Delta t \approx \frac{1}{2\pi}$$

This is the *same* uncertainty principle encountered in quantum mechanics: the particle's position and momentum are uncertain at the level of Planck's constant.

In information theory the uncertainty principle is related to the Shannon limit which states that it takes at least one Hz of bandwidth to send one bit of data in one second. But you can "beat" the uncertainty principle if signals are strong, and a more correct statement is

$$\Delta f \, \Delta t \approx \frac{1}{2\pi * SNR}$$

All of the modems which seem to violate the uncertainty principle and Shannon's limit do so by requiring stable signal and high SNR. One of the goals of HF modem development should be to figure out just which schemes can force bits through poor channels. Unfortunately, the FCC's amateur rules are very specific on what technology can be used. Much of the R&D work will require STA' s.

6. The Ionosphere as a filter: Normally amateurs like to think of the ionosphere as a mirror and that their signals "bounce" off the mirror. However this is an overly simplistic view. In reality the ionosphere is a plasma permeated by a magnetic field and it is the electrons in this plasma which are responsible for radio propagation. The electron density below ≈ 100 km altitude is insignificant, and it rises to a maximum at ≈ 350-400 km. As a signal travels through this region it is continually bent. At any point along this path the index of refraction n (ignoring the effects of the magnetic field) is given by

$$n^2 = 1 - \frac{f_n^2}{f^2}$$

there f is the signal frequency and f_n is called the plasma frequency. The square of plasma frequency f_n^2 is in turn proportional to the *in situ* electron density N_e.

If a signal is sent vertically into the ionosphere and returns to the sender, then the signal must have penetrated the ionosphere to the level where the signal frequency f equals the plasma frequency f_n, at which point the index of refraction decreased to zero. The highest frequency at which this can occur represents the peak plasma frequency in the ionosphere. This occurs in the F_2 region and is often called f_oF_2, the critical frequency at vertical incidence. Vertical incidence radars called ionosondes are use to measure the profiles of electron density up to f_oF_2. Typically f_oF_2 will range from l-2 MHz during solar minumum at nighttime to 12-14 MHz at solar maximum during the daytime. Occasionally (usually during the summer) patchy clouds with much higher electron densities (with f_n reaching as high as ≈ 50 MHz) will form in the E-region (100-120 km altitude) and give rise to sporadic-E propagation.

If the ionosphere is probed with a signal with $f > f_oF_2$, the vertical incidence signal will escape the ionosphere and travel off into space. However more oblique ray-paths will be gradually bent and may return to the earth. It is these signals which concern us here.

The gradual bending of signals on encountering an index which varies with height is easily seen visually on the desert or on a long stretch of paved road. The surface heating of the atmosphere in the first meter above the earth causes variations in the air's index of refraction which "reflect" (the proper word is *refract)* the sky as a visible mirage.. The shimmering seen in the mirage will have its analog in multipath distortion of radio waves.

If one looks at the equation for the index of refraction carefully, it will be noticed that the index of refraction n C 1. In Physics were taught that the index of refraction n is the ratio

$$n = c/v$$

where c is the speed of light and v is the speed of the signal inside the medium. If n c 1 then it would seem that v > c in contradiction to special relativity. To understand this paradox, we must understand the definition of the velocity of propagation in the medium. In the case of a plasma, the medium is dispersive, i.e. it's index of refraction changes with frequency. What happens is that the wavelength -- the distance between two points with the same phase --

shrinks in a plasma and the index of refraction could be expressed as:

$$n = \frac{\lambda_{plasma}}{\lambda_{vacuum}}$$

The velocity "measured" by the normal index of refraction is $v_{P,}$ the *phase* velocity which is given by

$$v_P = \frac{1}{2\pi}\frac{\Phi}{f}$$

The rate at which information is transmitted is called the *group* velocity v_g and it is given by

$$v_g = \frac{1}{2\pi}\frac{d\Phi}{df}$$

which is the slope of the variation of phase vs. frequency. The group velocity $v_g < c$ and it is this velocity that is governed by special relativity. *Einstein is happy!*

7. On the air with the ionosphere -- how lone is my baud? When we operate short paths near $f_o F_2$ (like on 80 meters) our signal undergoes many turns of phase shift as it is slowed down and turned around. All the little density variations in the ionosphere perturb the phase markedly and lead to the severe multipath distortion seen on 80M. Conversely when we operate long paths at oblique incidence angles (like on 10M or 15M long-haul links) the signal spends little time in the ionosphere so it has little distortion.

Back in 1980-81, this contrast was made very apparent in some BPSK tests we performed. Following the loss of the AMSAT Phase-3A spacecraft (when the Arianne launcher blew up), the Phase-3 telecommand team decided to try our 400 bits/sec Manchester-encoded PSK hardware on HF. The U.S. members of the team applied for and received an STA to use voice-bandwidth BPSK in the HF voice sub-bands (our foreign colleagues in Germany, Canada and New Zealand had much more liberal rules than we did). The first tests were between Ian, ZL1AOX and W3IWI on 10M; everything worked great. We exchanged digital data for hours on end even when signals were weak. Later tests between DJ4ZC, VE6SAT, WØPN and W3IWI on 20M worked pretty well also.

Then John, W1HDX near Boston and I tried 75M -- it was a disaster! Virtually no data could be exchanged despite good strong signals. We decided to try some simple bi-static radar tests to see why. With bi-static radar the transmitting and receiving stations are separated. Both John and I took our turn transmitting with the other listening. *[Lest the reader worry about the legality of radar transmissions on HF, what we actually sent was high speed CW with CW Identification. A MorseMatic keyer is a great pulse generator sending E E E E E at* 99 *WPM!]*

The receiving station generated a clock running at the same as the sender, and the received signals from an ordinary SSB receiver were observed on a scope synchronized to the independent receive clock. Lo and behold, the effects of multipath were discovered once more. The received signal would show an initial spike as the signals from the first path arrived. Then signals from other paths would arrive with random phases and the signal was all chopped up for a few msec. Then, after all the various wavefronts arrived the signal would stabilize. On the ≈ 600 km Massachusetts-to-Maryland path, it took 5-10 msec for the signal to stabilize. But our

400 BPS Manchester-encoded data had individual signaling elements only 1.25 msec long. Small wonder that our 75M PSK tests were unsuccessful.
RTTY and AMTQR operators on 80M and 40M know that 45 or 50 baud signals (about 20 msec per signaling element) usually works. But 110 baud ASCII (9.1 msec bits) is marginal. Very few have been successful with 300 baud Bell 103 packet transmission on 80M. Many military and commercial HF tests have similarly indicated that 50-75 baud is about the limit for HF paths shorter than ≈1000 km and frequencies below about twice the f_oF_2.

In the ≈ 1000-2000 km range covered by the 40M, 30M and 20M bands, clearly 300 baud does work -- witness the success of SKIPNET at moving large volumes of packet mail (albeit with low efficiency). But the AMTOR and RTTY stations running lower baud rates are able to move data when packet stations have difficulty. The strong evidence is that baud rates in the 100-200 range would be much more suitable when conditions are poor.

Since there are days when the ionosphere is good and days when it is horrid, an optimized strategy would be to have adaptive rates which change with conditions.

Coding and Data Rates: In the preceding discussion we talked in terms of BAUD rates -- the number of signaling elements that occur in one second. We now briefly mention some schemes which can (and will) be easily implemented in DSP hardware for testing. Which proves the most successful remains to be seen. In all of these schemes, each signaling element conveys more than one bit.

The first simple example is Quadrature Phase Shift Keying (QPSK). A carrier signal is phase modulated to one of four possible phases. Since there are four discrete states of the signal, then 2 bits of data can be encoded in each time slice. If the propagation medium is stable enough to permit 8 or 16 phase states (8PSK and 16PSK), then 3 or 4 bits could be conveyed. In addition to modulating the phase of the signals, QAM coding will be tested.

Commercial and military systems using multi-tone on/off keying (OOK) have been successful. We might design a scheme where 32 tones spaced 25 Hz apart are used, with one tone representing a bit. The 25 Hz channel spacing and the uncertainty principle discussed earlier mean that we can key the tones at 25 baud (40 msec signaling elements) and occupy about 850 Hz of spectrum. Although this is a 25 baud system, it conveys 800 bits/sec of data. Instead of using all 32 bits for data, we might choose to encode the data with an error correcting polynomial; depending on the degree of protection desired this might reduce the delivered data rate to ≈600-700 bits/sec. This multi-tone OOK approach, if carried to its limit, places severe limits on the dynamic range in transmitters; the peak-to-average power requirements are quite large. Since amateurs are notorious for "cranking up the wick" and reducing the peak-to-average ratio, on-the-air tests will be needed to see how gracefully multi-tone OOK looses performance and how "unfriendly" it is to users on nearby frequencies.

In the early days of packet radio, error correction was rejected because at the time the CPU chips in the VADG (8085), TNC-1 (6809) and TNC-2 (Z-80) lacked the "horsepower". The AMTQR protocols placed less stringent requirements on their embedded processors so FEC was included. Now the price of smart silicon has come down so the early decision should be revisited, especially for HF applications.