                VOICE FILE INTERCHANGE PROTOCOL (VFIP)

STATUS OF THIS MEMO

   This memo describes a proposed voice file interchange format for use
   in the ARPA-Internet community. Suggestions for improvement are
   encouraged. Distribution of this memo is unlimited.

1. INTRODUCTION

   The purpose of the Voice File Interchange Protocol (VFIP) is to
   permit the interchange of various types of speech files between
   different systems. Currently, there are many different types of
   voice implementations, but no specific standard has been set with an
   eye towards compatability between these systems. With the increasing
   interest and development of voice, specifically in Multimedia Mail,
   there is an increased need to include standardized speech into a
   common data structure.

   The Voice File Interchange Protocol defines a header to describe the
   voice data. The 18-byte header contains the identifier, the header
   version number, the header length, a DTMF mask for Touch-Tones, the
   recording rate in bits per second, the total time in deci-seconds
   (tenths of a second), and the encoding/recording method (see
   Figure 1).

2. THE VOICE FILE INTERCHANGE PROTOCOL HEADER

   The Voice File Interchange Protocol header is organized as follows:

   2.1 The Header Version Number

      The version number is 1-byte. This first version is number one.

   2.2 The Header Length

      The length is a 1-byte field indicating the length of the entire
      header in bytes. For this first version, the length is
      18 (bytes).

2.3 The DTMF Mask

   This field describes what is known about DTMF Touch-Tones in the
   data. The field consists of a 16 flag bits which indicate what is
   known about particular DTMF tones. The 16 possible DTMF tones, in
   order, are: 0 1 2 3 4 5 6 7 8 9 # * A B C D. The low order bit
   of the field is tone 0.

   A 1-bit signifies that the corresponding tone is guaranted NOT to
   be in the speech file. A 0-bit signifies that it may or may not
   be in the speech file. Therefore, a field of 16 zeros denotes
   that nothing is known about the tones. A field of 16 ones denotes
   that there are no tones in the file.

2.4 Recording Rate

   The recording rate is a 32-bit field and is the approximate rate
   in bits/second of the method used to record the speech. For
   variable rate methods, this may be very approximate.

2.5 Total Time

   A 32-bit number indicating the total time of the recording in
   deci-seconds. For example, 600 indicates 1 minute of speech.

2.6 Methods of Encoding/Recording

   This 6-byte ASCII field indicates the method of
   encoding/recording. Names shorter than six characters are padded
   out to the right with blanks (the ASCII space character, code 32
   decimal). For comparisons, the names are case insensitive.

   Some known methods of Encoding/Recording are:

     TI - The Texas Instruments card for the IBM PC [5].

     IBM - PC Voice Communications Options.

     NVP-1 and NVP-2 - Network Voice Protocol [1,2].

     COMPUT - Computalker card for the IBM PC [4].

3. SUMMARY

   This 18-byte header will permit interchange of speech files between
   different systems, as well as facilitate automatic conversion between
   formats. The header does not have to be prepended to the speech file
   proper; it may be in the form of a separate associated file, if that
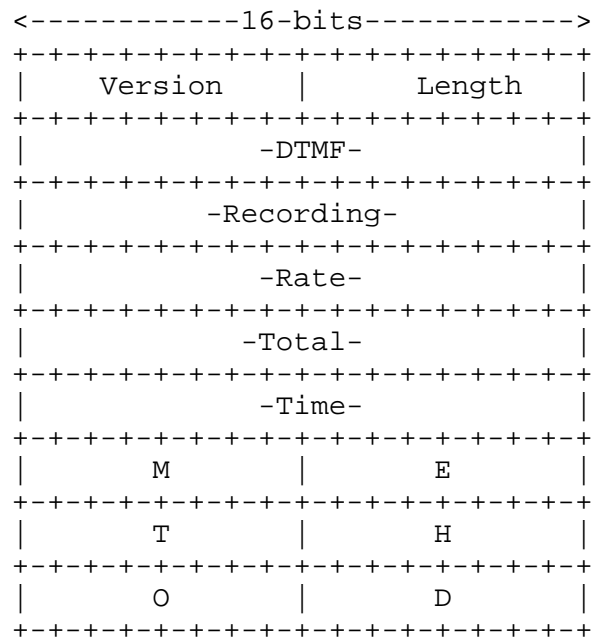   is more convenient.

```
                    <------------16-bits------------>
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |    Version    |    Length     |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |             -DTMF-            |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |          -Recording-          |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |             -Rate-            |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |            -Total-            |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |             -Time-            |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |       M       |       E       |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |       T       |       H       |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |       O       |       D       |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

                          Figure 1

4. EXAMPLES

    Example 1 is for one minute of 2400 bps NVP-2 speech. Nothing is
    known about DTMF tones in the data.

```
              <------------16-bits------------>
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |      1        |      18        |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |               0               |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |                               |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |             2400              |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |                               |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |             600               |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |      N        |      V         |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |      P        |      -         |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |      2        |     <sp>       |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

                             Example 1

Example 2 shows the header for 10 seconds of 1200 bps TI speech, with
none of the DTMF tone 0-9 in the data, but no information about
tones *, #, A-D.

```
              <------------16-bits------------>
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |     1        |       18        |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |              1023              |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |                                |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |              1200              |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |                                |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |              100               |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |     T        |       I         |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |    <sp>      |      <sp>        |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              |    <sp>      |      <sp>        |
              +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

                         Example 2

REFERENCES

   [1]   Cohen, Danny, "Specifications for the Network Voice Protocol
         (NVP)", RFC 741 (NIC 42444), USC/Information Sciences Institute,
         January 1976.

   [2]   Cohen, Danny, "A Network Voice Protocol (NVP-II)",
         USC/Information Sciences Institute, April 1981.

   [3]   O'Leary, G. C., "Local Access Area Facilities for Packet Voice",
         MIT/LL, October 1980.

   [4]   Computalker, "Compu Phone for the IBM PC/XT", Santa Monica,
         California, August 1985.

   [5]   Texas Instruments, Inc., "The TI Speech Application Tool Kit
         Guide", TI Part #2232384-1, May 1985.