Open access • Journal Article • DOI:10.1109/25.875264

# Mobile audio-visual terminal: system design and subjective testing in DECT and UMTS networks — Source link

D. Gill, John Cosmas, Alan Pearmain

**Institutions:** ERA Technology Ltd, Queen Mary University of London

**Published on:** 01 Jul 2000 - IEEE Transactions on Vehicular Technology (IEEE)

**Topics:** Speech coding, Audio signal, Code-excited linear prediction, Transform coding and Linear predictive coding

# Mobile Audio–Visual Terminal: System Design and Subjective Testing in DECT and UMTS Networks

David Gill, John Paul Cosmas, *Member, IEEE*, and Alan Pearmain, *Member, IEEE*

***Abstract*****—It is anticipated that there will shortly be a requirement for multimedia terminals that operate via mobile communications systems. This paper presents a functional specification for such a terminal operating at 32 kb/s in a digital European cordless telecommunications (DECT) and universal mobile telecommunications system (UMTS) radio network. A terminal has been built, based on a PC with digital signal processor (DSP) boards for audio and video coding and decoding. Speech coding is by a phonetically driven code-excited linear prediction (CELP) speech coder and video coding by a block-oriented hybrid discrete cosine transform (DCT) coder. Separate channel coding is provided for the audio and video data. The paper describes the techniques used for audio and video coding, channel coding, and synchronization. Methods of subjective testing in a DECT network and in a UMTS network are also described. These consisted of subjective tests of first impressions of the mobile audio–visual terminal (MAVT) quality, interactive tests, and the completion of an exit questionnaire. The test results showed that the quality of the audio was sufficiently good for comprehension and the video was sufficiently good for following and repeating simple mechanical tasks. However, the quality of the MAVT was not good enough for general use where high-quality audio and video was needed, especially when transmission was in a noisy radio environment.**

## I. Introduction

During the 1980s, there were dramatic changes in the area of mobile communications. This was especially true in the business world. At first, there were only mobile telephones in taxis and tone-only pagers in hospitals. It quickly became commonplace to hear a pager, use a domestic cordless telephone, or use a first-generation mobile telephone. The 1990s are currently welcoming second-generation mobile systems such as global system for mobile telecommunications (GSM), digital European cordless telecommunications (DECT), and CT2/CT3 (cordless technology). This is instigating a proliferation of technologies and services. The next step to be taken toward the turn of the century will be the introduction of a global mobile system called universal mobile telecommunications system (UMTS). This will probably merge paging, cordless telephones, mobile terrestrial, and mobile satellite standards into a single unified standard.

Video coding at very low bit rates (VLBRs), in the range of a few tens of kilobits per second, is becoming very attractive for a number of new applications, such as mobile video communication, video telephony on the public switched telephone network (PSTN), multimedia electronic mail, and remote sensing, and for interactive data bases. The ability to transport compressed audio and video over mobile links will open up new areas of opportunity for services not yet commercially developed and provide the incentive to migrate from GSM to UMTS networks. Communications can be provided rapidly where there is an urgent need, in the form of mobile terminals, without the costly overhead of cable provision. The area of security surveillance could be greatly enhanced as mobile security systems could be set up very quickly whenever and wherever required. The ability to send audio and video to and from mobile units could be of great benefit to the emergency services.

The transmission of uncompressed video is very expensive. A single broadcast television channel requires in excess of 100 Mb/s. In December 1990, after five years of international cooperation, the CCITT recommendation H.261 for audio–visual transmission of video telephony and video conference, at bit rates between 64 kb/s and 2 Mb/s, was adopted. H.261 provided lower transmission costs and a unified standard giving global compatibility, which is extremely important for the expansion of audio–visual services. The problem with H.261 is that it was developed for use over fixed links and is not very well suited for use over mobile channels which may be prone to poor channel error performance. The large compression ratios achieved by H.261 are obtained by extracting much of the redundancy from the input video stream. This leaves the remaining data stream very vulnerable to errors. If an error finds its way into the data stream, the effect on the video image would be seen easily and it could remain on screen for several seconds. References [1] and [2] describe an experiment in which an H.261 data stream was carried over a DECT channel. This work showed that new audio and video compression algorithms need to be researched, designed, and constructed. The high-error rate possible for a mobile radio link requires that techniques for improving the error resilience of the coding scheme be investigated.

References [3] and [4] outline the European plan of action for research and development in this area of mobile communications. The development of new services for mobile communications is of utmost importance for the success of UMTS. This multimedia market will only succeed if users are sure that they will be able to share their multimedia information across different platforms in an easy and seamless way. In the context of these developments, a demonstrator mobile audio–visual terminal (MAVT) has been designed and constructed to demonstrate real-time moving video and audio over a low bit rate mobile radio channel.

Manuscript received March 31, 1998.

D. Gill is with the Communication Systems Division, ERA Technology Limited, Leatherhead Surrey, U.K.

J. P. Cosmas is with the Department of Electronic and Computer Engineering, Brunel University, Middlesex UB8 3PH, U.K.

A. Pearmain is with the Department of Electronic Engineering, Queen Mary and Westfield College, University of London, London E1 4NS, U.K.

Publisher Item Identifier S 0018-9545(00)03695-1.

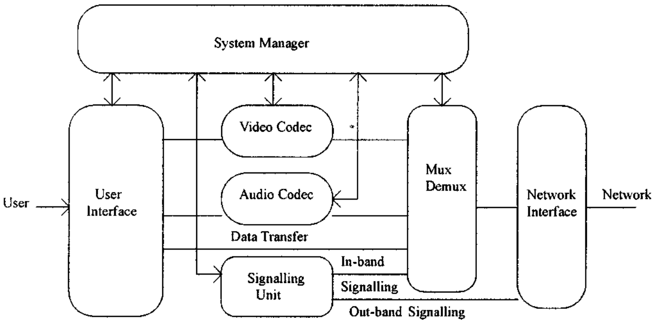


Fig. 1. General structure of an MAVT.

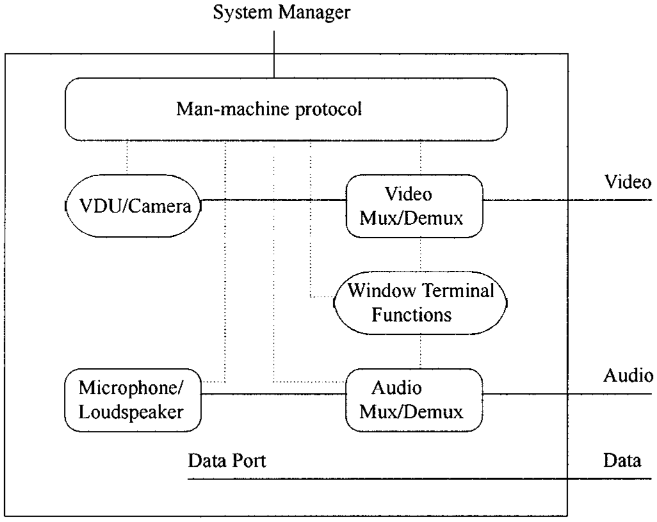


Fig. 2. User interface block diagram.

## II. Structure of MAVT

The CCITT has standardized audio–visual terminals (with data transfer capability) for $p$*64-kb/s digital connections using the recommendations H.320 and H.261. The overall block diagram of an MAVT fits inside the general framework of a CCITT H.320 terminal. The characteristics of mobile networks and the introduction of new services naturally affect the contents of the different functional blocks. Fig. 1 shows the general block diagram for an MAVT. It is also described in [5] and [6].

### *A. System Manager*

The system manager forms the central intelligence of the terminal. All communication between different functional blocks takes place via the system manager. All decisions with respect to the terminal states are taken by the system manager. This avoids the daunting complexity of a distributed system.

### *B. User Interface*

Fig. 2 can be made more symmetric with regard to the user-network duality. A multiplexer function and a man–machine protocol unit are functionally part of the user interface. Icons and touch-screen buttons can be used for user input. Warning bells and stored messages can be used as feedback to the user. Computer-generated images can be routed to the user instead of the camera output. Therefore, the input to the video display unit and the loudspeaker input is subject to multiplexing functions. The user interface must contain low-level routines to visualize windows, menus, and icons. In this respect, the user interface must make available a set of MS-Windows or X-Windows functions or an equivalent thereof. On top of this basic layer of the user interface is the man–machine protocol. The man–machine protocol allows the user to select the desired audio–visual service, dial numbers, etc. Fig. 2 gives an overview of the user interface block diagram.

### *C. Radio Interface*

The radio system is provided by a commercial DECT product, the Siemens Gigaset. A small interface module has been designed to connect the DECT handset to a digital signal processor (DSP) board. The MAVT delivers a fixed data stream rate of 32 kb/s. Stuffing bits are inserted if necessary. The data stream includes 8-kb/s audio data, 23.2-kb/s video data, and 0.8-kb/s control data. The control data consists of a frame alignment signal (FAS) and a bit allocation signal (BAS). The FAS includes a Willard word which is used for synchronization. The BAS word is protected using a BCH code. The whole structure is similar to the H.221 standard introduced for ISDN video phones. A DECT frame lasts 10 ms. The user information of a DECT frame consists of 320 b, arranged in 40 octets.

The DECT radio link consists of a Gaussian minimum shift key (GMSK) modulator that generates a signal which is transmitted in a Rayleigh fading multipath propagation radio channel with additive white Gaussian noise (AWGN). A mobile transmission channel is prone to more severe impairments than a stationary channel due to the effects of multiple scatter, frequency and/or time dispersion, shadowing, path loss, etc. The effect of the multipath propagation is to randomly attenuate the transmitted multipath signal as a function of distance traveled or time (given the velocity of the mobile). For a fixed noise level at the receiver this has the effect of randomly changing the signal-to-noise ratio (SNR) at the receiver. A DECT slot occurs every 10 ms, and it takes as long as ten DECT slots for a mobile moving at 3.6 km/h to move from a destructive fading situation to a constructive fading situation. This means that errors will be bursty. The received signal is differentially demodulated by differencing the received signal phase of the previous and present sample. A positive phase change represents a symbol 1 and a negative phase change a symbol $-1$. A full analysis of the DECT radio link is made in [7] where in order to decorrelate the bursty errors nine DECT frames are interleaved into a compound frame. The network interface is specific for the particular kind of network being used.

With respect to the signaling information a distinction can be made between user-network signaling and out-of-band end-to-end signaling. For instance, with DECT, at call setup with the aid of service attributes a terminal can indicate "H.261 video telephony." However, in general the network will not allow for all necessary end-to-end signaling capacity. Therefore, in-band signaling is necessary. The receiving terminal is instructed by means of remote commands on how to demultiplex and decode the signal using remote commands.

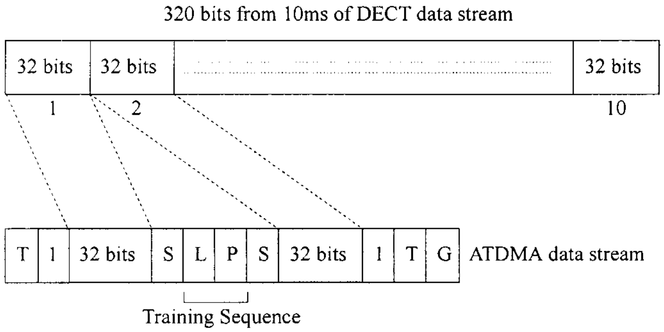


Fig. 3. DECT to ATDMA conversion.

## D. The ATDMA Interface

In this setup, the DECT interface module is connected to an ATDMA radio network. The new interface board converts the MAVT data into a V11 interface format which can then be fed directly into the ATDMA real-time testbed.

In 10 ms of DECT data capture time, there will be 320 b of MAVT data available. This data will cross the channel coder without being protected or interleaved. The task of the channel coder is simply reduced to adding stuffing bits in order to fit the format of the burst payload to be sent to the frame builder. The channel decoder then only has to remove these stuffing bits.

The ATDMA burst, as shown in Fig. 3, is composed of a training sequence in the center (L+P) used for channel estimation and synchronization purposes: tails symbols (T) at the end of the burst used mainly for equalization purposes; in-burst signaling (S) to transport signaling information; and a guard time (G) between successive bursts and data payload for information transmission. The reference documents [8]–[10] provide a full description of the ATDMA-MAVT terminal interconnection.

A modified version of the system manager has been written which redirects the outgoing MAVT data to the C40 processor on the new interface board, translating the data into a data stream that can be handled by the ATDMA interface. Another software path returns data received by the ATDMA interface to the system manager. The ATDMA interface is activated by a user option which has been added to the user interface.

## E. Multiplexer

In general, there is a need to communicate to other terminals what kind of audio–visual information is sent (multiplexing format, codec parameters used) and what kind of audio–visual information can be received, e.g., the capacity to receive user data. This is the subject of an end-to-end protocol (H.242 in the case of H.320 terminals). The signaling unit accommodates the end-to-end protocol as well as the user-network protocol. The end-to-end protocol can use both in-band and out-of-band signaling. The in-band signaling data and/or encoded audio, encoded video, and user data arefed into a multiplexer which enforces a certain framing structure onto the channel data. Also, the signaling unit provides error correction/detection within its own data substream. Forward error correction of the user data is optional. Video text, database retrieval, and file transfer are applications that fall into the category of user data. By means of in-band signaling, the particular application can be indicated and the appropriate codec can be activated.

The task of the multiplexer is to gather information from the various modules and combine them into a format that can be passed to the network, in this case the DECT radio system. This information includes both error protected audio and video data and control codes. The multiplexer is activated by a call (usually a hardware interrupt) from the network.

Data is generated by the audio codec and the video codec. Control information is generated by the user interface and the system manager. The users select their requirements on the PC via the user interface. The system manager accesses this data and passes the appropriate commands to the codecs. The system manager can then read the required data from the codecs. This data is then put into an intermediate frame buffer. A correct control word and frame synchronization word is then added. Finally, the data is transferred to an output buffer where the data will be sent to the DECT radio system.

The task of the demultiplexer is to read the incoming data from the network, in this case the DECT radio system. The demultiplexer splits the incoming data into its separate components and passes it to the proper decoders. The demultiplexer is activated by a call (a hardware interrupt) from the network. In most cases, when the demultiplexer is activated, it has to locate a frame synchronization word of some kind. Once the frame word has been found, then the system is deemed to be synchronized. Assuming that the synchronization has been achieved, data can then be extracted from the buffer and sent to the relevant decoders.

## F. Video Codec

The video codec realized in the MAVT DECT demonstrator [11], [12] is of the block-oriented hybrid discrete cosine transform (DCT) type. The incoming QCIF images are divided into $22 \times 18$ nonoverlapping blocks of size $N \times N$ with $N = 8$ for the luminance and $N = 4$ for the chrominances. Temporal redundancies are reduced by motion compensation with half pel resolution. Spatial redundancies are reduced using DCT coding both for interupdate and intraupdate. The block diagram of the encoder is depicted in Fig. 4. A special channel coding comprising unequal error protection is used to adapt the system to error prone environments such as the DECT channel. The MAVT codec is not compatible with existing or proposed standards such as H.261 [13] and H.263 [14]. The image size of the input and output sequences is a quarter common intermediate format (QCIF) in $4:2:0$ color format according to Table I.

To avoid image degradation by repeated filtering, the format conversion from QCIF to CIF is performed only once for the complete first image and once per DCT update block. To start the transmission of a sequence of images, the first image is built up by coding the mean values of all blocks. The mean value of a block of size $N \times N$ is quantized with 6 b and linearly predicted considering the last transmitted block (at the beginning of a new line of blocks, the predicted mean value is 128). The prediction error is fed to a nonadaptive arithmetic coder [15], [16]. To avoid very long and thus very error sensitive blocks of arithmetically coded data, the image startup information is split

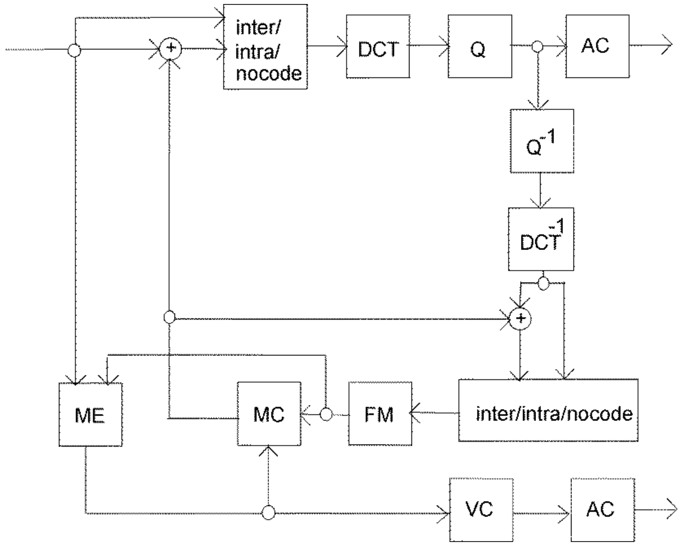


Fig. 4. Block diagram of the MAVT block-based source encoder. ME: motion estimation coding, MC: motion compensation of DCT coefficients, FM: frame memory field coding, AC: arithmetic, Q: quantization, and VC: vector.

TABLE I
IMAGE FORMATS

| Image Component | QCIF | CIF |
|---|---|---|
| Luminance Y | 176 pels x 144 lines | 256 pels x 288 lines |
| Chrominance U=B-Y | 88 pels x 72 lines | 176 pels x 144 lines |
| Chrominance V=R-Y | 88 pels x 72 lines | 176 pels x 144 lines |

into five independent segments. The first three segments code the lines from 0 to 47, 48 to 95, and 96 to 143 of the luminance, respectively. The next segment contains the chrominance B-Y and the last the R-Y information.

Motion estimation is performed on 8 × 8 blocks with half pel resolution, based on the luminance of the current original image (QCIF) and the previous reconstructed image in CIF format. After a three-step block matching minimizing the square error [17], the vector field obtained is refined by the so-called Gibbs smoothing. Since a full search is not feasible in the current hardware implementation, three-step block matching is used which is much cheaper in number of computations.

Gibbs smoothing of the displacement vector field reduces the number of bits necessary for coding the motion vectors and leads to a more natural description of the motion [18]. Usually, the SNR of the motion-compensated image is slightly lower after smoothing, but this is more than compensated by the bits saved for vector coding that improve the update.

The displacement vectors are coded by a combination of quadtree and linear predictive coding. Both components of a vector (vertical and horizontal component) are coded with one symbol, not separately.

Quadtrees are used to exploit areas with constant motion [19]. Whenever all motion vectors in one group are identical, only one vector is coded for this group. The structure of the quadtrees is coded from the upper left to the lower right corner of the image.

In quadtree areas with constant vectors, only the upper left vector is coded. All vectors to be transmitted are predicted from preceding vectors. The prediction error (the difference between vector and its prediction) is coded arithmetically.

The motion-compensated image in CIF resolution is constructed by shifting blocks from the previous reconstructed CIF image according to their motion vectors. The motion vectors for the chrominances are derived by dividing the luminance vector by two and truncating the result to an integer value. The input for the DCT update consists of the motion-compensated image in QCIF resolution and the current original QCIF image. The motion vectors are derived by subsampling. Uncoded blocks are copied to the current reconstructed CIF image. Coded blocks are first interpolated to CIF resolution.

After motion compensation, there are, for each block independently, three possible modes for the further processing.

1) No code: the previous block is directly copied into the current reconstructed image.
2) Interupdate: the difference between the motion-compensated and the original block is DCT coded.
3) Intraupdate: the original block is DCT coded.

First, the inter/intradecision is taken. Afterwards, the code/no code decision is taken on the basis of the number of available bits. All mode decisions for a block are valid for the 8 × 8 luminance block as well as for the two corresponding 4 × 4 chrominance blocks. The DCT coefficients are then quantized. For interblocks, quantization resolution is identical for all coefficients and for intrablocks quantization resolution for the dc coefficient, and the other ac coefficients are different.

Before coding, the quantized DCT coefficients are zigzag scanned to yield a one-dimensional (1-D) vector. In the case of intrablocks, the dc coefficient is directly coded using 8 b for luminance and 7 b for chrominance blocks. The ac coefficients of intrablocks and all coefficients of interblocks are run length coded, i.e., for each nonzero, the number of zeros before this coefficient (run length), and the coefficient itself is arithmetically coded in two different streams. The first stream consists of all run lengths occurring in one image, with the separator symbol end of block (EOB) after each block and end of string (EOS) after the whole image. This stream is fed into a nonadaptive 66-symbol (run lengths 0–63 plus EOB plus EOS) arithmetic coder. The second stream consists of the corresponding nonzero DCT coefficients. Only an EOS symbol has to be appended, since EOB's would be redundant.

The inter/intradecision is performed similarly to H.261 RM8 [20], [21], but on an 8 × 8 block basis. For the selection of coded blocks, all coded blocks, inter and intra, use the same quantizer step size. Beginning with the block with the highest motion-compensation error, the blocks are recursively marked as to be coded as long as bits are available. The selection of the coded blocks is based on their squared error.

For coding of block positions, the position $(i; j)$ of each DCT-coded block is coded by spending 9 b for its index. Additionally, there is a stream comprising one bit per coded block representing the inter/intradecision and for each intrablock the three dc coefficients (luminance +2 chrominance).

There are several bit streams generated by the hybrid coder that represent different kinds of information.

1) First image: mean values of all blocks [arithmetic (AR) coded].
2) Motion compensation: quadtree.
3) Motion compensation: motion vectors (AR-coded).
4) DCT update: positions of update blocks.
5) DCT update: quantizer, inter/intradecisions and dc coefficients.
6) DCT update: ac coefficients and run lengths (AR-coded).

### G. Video Channel Coding

In general, individual symbols or symbol groups in the source data stream, as for low bit rate video and audio coding, exhibit different sensitivities. This characteristic suggests an unequal error protection, and therefore an efficient partitioning of the additional redundancy is needed. A further step would imply the common design of source and channel coder. An unequal error protection can be realized in principle through block codes. The application of convolutional codes offers some crucial advantages, since a "soft-decision" maximum likelihood (ML) decoded signal with channel significant information through application of the Viterbi algorithm (VA) yields a significant gain with respect to the $E_b/N_o$ ratio. Furthermore, only one coder and decoder is needed. Unequal error protection is achieved through puncturing a so-called mother code of rate $1/n$ and memory $m$ periodically with period $P$. The underlying puncturing matrix must fulfill the rate-compatible restriction in order to guarantee an average free distance between transitions in case of a dynamic redundancy allocation. Rate-compatible punctured convolutional (RCPC) codes represent an efficient method for unequal error protection matched to the characteristics of the underlying source coder [22], [23].

The design of the proper RCPC code involves knowledge of the symbol sensitivity of the source coder and the protection required in the case of the underlying transmission system (modulation, propagation channel). The sensitivity of a special source coding algorithm can be determined through introduction of error in the coded source data frame and application of a relevant criteria for measurement of the effects after the source decoder. A widely used objective criteria is the SNR after decoding, which includes distortion due to quantization and transmission errors. In case of a sufficient word length, the corresponding error terms are uncorrelated, so that the error energy due to transmission errors alone can be isolated. Since a proper error criterion must be perceptually relevant, a long-weighted noise-to-signal ratio is applied in the field of speech and video coding. In this case, segments with weak energy are emphasized with respect to the conventional SNR. In general, one must consider additional subjective criteria for proper analysis of the source coder sensitivity.

Considering the sensitivity information acquired, a finite classification of symbols with similar error sensitivity in a finite number of groups is carried out. On the average, each group should contribute the same amount after error protection to the overall noise due to transmission errors, which should be minimized. Based on this, the bit error rate required for decoding can be determined. The correction capability or code rate to achieve the necessary transmission quality can be obtained through computer simulation of the transmission system.

The MAVT video channel encoder has a constraint length of seven, uses the three generator polynomials $0 \times 5b$, $0 \times 79$, and $0 \times 65$, generates a mother code rate of 1/3, and uses rate-compatible puncturing pattern coding with a resulting code rate from 1/3 to 8/9 and a periodicity of eight. Before sending data to the transmission channel, the channel encoded data stream is passed to the interleaver [24].

### H. Audio Coding

In order to satisfy the above requirements this section presents the processing applied to the speech signal in the MAVT demonstrator which includes speech enhancement (echo cancellation and noise reduction), so that the demonstrator can be used in hands-free mode: source coding at a bit rate of 4.6 kb/s, channel coding (including interleaving) to reach a net bit rate of 8 kb/s, and the use of voice activity detection so that the audio bit rate can be allocated to video coding when there is no voice activity. These processes are:

1) preprocessing of the speech signal to remove the surrounding noise and echo picked up by a hands-free terminal;
2) source encoding/decoding at a low bit rate (4.6 kb/s) through the low-complexity RP-CELP algorithm;
3) channel encoding/decoding, including block interleaving, for protection against transmission errors; the net bit rate is then 8 kb/s.

In audio communication, the use of a hands-free receiver/transmitter results in two kinds of disturbing signals that interfere with listening comfort or even intelligibility of speech:

1) surrounding noise, picked up by the transmitting hands-free microphone;
2) echo signals, due to the coupling between the loudspeaker and the microphone of the hands-free system.

Therefore, the hands-free preprocessor must include an echo cancellation system and a noise reducer to solve both problems. These two processors are cascaded, the echo canceler being first. The final enhanced speech is then directly fed to the speech source coder. A detailed description of both algorithms is given in [25] and [26]. Echo cancellation is based on a digital adaptive filtering, called adaptive echo canceler (AEC), using the normalized-least mean squares (N-LMS) algorithm [27]. The AEC consists of subtracting from the actual echo an estimated form at the microphone output. This estimated echo is computed from the far-end speaker's signals, received at the loudspeaker input and processed by a filter which models the impulse response of the echo paths. In order to track the echo variations, this filter is time adaptive, updated by the N-LMS algorithm. Its adaptation is frozen when the near-end speaker talks (i.e., double-talk situations). The resulting signal, corresponding to the error between the actual echo and its estimate, is sent to the noise reduction stage of the hands-free preprocessing.

Noise reduction, based on magnitude spectral subtraction, is applied after echo canceling. Echo-canceled speech signals, which still include ambient noise, are transformed in the spectral domain by the fast Fourier transform (FFT) and processed by a magnitude spectral subtraction filter [28], [29]. In order to keep the speech signal undistorted, spectral flooring is applied on the output filter signal. Then the enhanced speech signal is obtained by applying an inverse FFT combined with an overlap-add operation and is sent to the far-end network. The incoming signals are processed by blocks or frames with a length of 240 samples (30 ms at 8 KHz) although Fourier analysis is performed on overlapping frames of 256 samples (32 ms).

This noise reduction system gives excellent results on stationary noise (such as car noise). However, in certain applications of the MAVT where the noise does not have the stationary property (e.g., a construction site) this process should not be applied.

The speech codec is a regular pulse code-excited linear prediction (RP-CELP) system [30] based on the CELP technique, which uses the properties of speech as an audio signal. The codec operates at a net bit rate of 4.6 kb/s. The speech signal is processed on a frame-by-frame basis, with a frame length of 30 ms (240 samples at 8 KHz).

The analysis-by-synthesis scheme consists of three parts. First, a tenth-order linear prediction filter is computed, quantized, and applied to remove the short-term correlation. Then the encoder processes the four successive subframes of 60 residual samples. Long-term prediction (LTP) analysis is performed using fractional delays and a closed-loop procedure. The last stage, or regular pulse code (RPC) stage, involves a structured codebook which is made up of four different binary regular pulse subcodebooks. The best codeword is determined with respect to a convenient perceptually weighted mean-squared error criterion that does not require an exhaustive search. After each step (LTP and RPC stages), a local decoding procedure is applied to compute an estimated residual signal, which completes the feedback loop of the analysis-by-synthesis scheme.

The coding process results in a set of quantized parameters (linear prediction, LTP lag and gain, and codebook index and gain), which is transformed in a block of 140 b, ordered by decreasing subjective importance. This block of 140 b is then delivered to the channel coder.

The speech decoder receives from the channel decoder an equivalent block of 140 b, plus one information bit known as the bad frame indicator (BFI). This last information is used to replace the corrupted parameters of the current frame by extrapolating the ones received in previous valid frames. The decoded parameters, either reconstituted from the 140 input bits or extrapolated from previous ones, are then used to reconstruct a frame of 240 synthetic speech samples. For each subframe, the excitation codeword, scaled by its gain factor, is computed and run through the long-term synthesis filter. It yields a reconstructed residual signal which is finally run through the short-term synthesis filter, thus providing the reconstructed speech frame.

A voice activity detector is included in the noise reduction unit [31], [32]. This information is used to distinguish between speech and noise in the signal received, for estimation of the noise parameters. But another possible application for this information is the decision to transmit or not the audio parameters. If no speech information is included in the audio signal (pure noise = voice activity indicator low), then the speech bit rate may profitably be used by the video unit.

However, the noisy environmental conditions when noise reduction cannot be applied (see that section) make it impossible to stop transmitting the audio signal altogether when no voice activity is detected. In such conditions, it is very uncomfortable for the listener to have moments of noisy speech alternate with moments of utter silence.

It is necessary to transmit information about the ambient noise (called comfort noise) even if no speech information is present. This transmission consists of isolated frames, repeated at regular intervals. Due to channel interleaving, the noise will be transmitted in blocks of three audio frames (90 ms) and updated every 12 frames (360 ms).

Last, the transmission of the speech signal must resume immediately when voice activity is detected, which calls for permanent anticipation of two audio frames (due to block interleaving) between VAD and information transmission. Of course, the information as to whether audio data is being transmitted has to be provided to the MAVT System Manager and transmitted through the channel.

### *I. Audio Channel Coding*

The channel coder receives from the speech coder a block of 140 b, ordered by decreasing subjective importance. The first ones are more sensitive to errors, whereas the last ones are more robust. Therefore, the protection scheme considers three different importance classes. In order to detect errors on the most significant bits of the speech frame, a four bit CRC is performed over class I. To protect the largest number of bits, a convolutional code with a memory of 4, rate 1/2, and constraint length of 5 is applied to classes I and II of the frame, whereas the third class contains unprotected bits.

The output of this process is a block of 720 b for 90 ms of speech data (three audio frames), which corresponds to the required bit rate of 8 kb/s. Interleaving is then applied, the 720 b being spread over nine DECT bursts of 80 audio bits each.

In the decoding part, the 720 b of the three speech frames are deinterleaved, to obtain the reordering of the $3 * 240$ b. Convolutional decoding is performed using the Viterbi algorithm. The result of CRC decoding is used to detect errors on the most significant bits. If an error is detected in a frame, the erroneous frame is not transmitted to the speech decoder and the BFI is set to one. In addition, BFI is set to one when the quality factor, which is computed after reencoding of the decoded frame, is lower than a certain predefined threshold. In other cases, BFI is set to zero. The 140 decoded bits plus the BFI are transmitted to the speech decoder, within a frame of 141 b, at the beginning of which the indicator has been set. A complete description of the channel codec (speech and video) for DECT is given in [33].

## III. DECT AND UMTS TESTS

The MAVT field tests use at most two MAVT's, connected by either a DECT or UMTS radio link as a transmission network. Additional test equipment is required for lighting, video recording, audio recording, etc. A full set of tests would include typical indoor and outdoor scenarios and a subset of possible mobile audio–visual applications.

The subjective field tests should comprise a task and an exit questionnaire. Each task will be divided into approximately 1-h sessions to avoid fatigue in the subjects. It is recommended that they be allowed a 10-min break with refreshments before the tasks are resumed. The tests are administered by someone who can put the subjects at ease, explain the proceedings and record/collect the appropriate results. Naive users do not want to be concerned with the technical details or optimization of an MAVT before or during use. Most of the optimization should be done automatically, and during the tests the administrator is asked to identify any difficulty that the users experienced.

### A. Video Subjective Tests

Possible subjective tests of the system are viewer impressions, viewer tests, interactive tests, and communications tests.

*1) Viewer Impressions:* In this subjective test, the subject evaluates the video quality of the MAVT. The subjects are asked to comment on various aspects of the video image. Does the video resolution of the image give a good likeness of the subject matter? Is the frame rate of the video display fast enough to give a good indication of the motion of the subject? Does the color of the video give a true representation of the subject?

For this test, it is necessary for the subjects to modify the video parameters (brightness, contrast, and saturation) before the start of the test. Each subject is able to choose what they consider to be the optimum settings for the video parameters. These settings are recorded for each subject.

*2) Viewer Tests:* The idea of the viewer test is to ensure that the user can readily identify objects portrayed on an MAVT screen. This task requires only one subject. The subject is given a fixed amount of time to become familiar with 20 solid objects. Subsequently, the subject is asked to identify them one at a time using the visual aspects of the MAVT only. The subject can ask the operator at the remote MAVT to change the orientation of the object.

Before the task commences the subject is asked if he/she would like to change any visual preferences. Each physical object is put in front of the remote camera one at a time for identification. The operator is not limited to using each object once. The operator at the remote MAVT moves the test object according to the subject's requests. If the subject is unable to identify the object after some time, he/she may freeze the video image. If the subject still cannot identify the image, he/she may record and view several still images from different angles as a last resort.

The remote MAVT has its audio input disabled (e.g., disconnecting the microphone) so that the operator cannot offer any audible advice or feedback. This is in the event that the subject asks the operator a question which he/she should not answer. The administrator records:

- facilities that the subject used;
- objects displayed;
- what identification the subject made of the objects;
- number of correct identifications;
- mode that was used to identify the object;
- time taken to identify each object;
- total time taken;
- time when the experiment started;
- any other useful observations/comments.

*3) Interactive Tests:* The subjective interactive test is to ensure that it is possible for two users to interact with each other coherently over an MAVT link. The task chosen for this test is a simple model building exercise. The model building is performed by two subjects. The aim is for one of them to build a model out of LEGO while the other watches, over the video link of the MAVT, and attempts the construction. Each subject is given an identical set of LEGO pieces and asked to build a model from a list provided. Each subject takes turns to build their chosen model. During this task the subjects are not allowed to talk to each other. The builder of the original model is asked to go slowly and take his/her time. The administrator records:

- what model they built;
- start and ending times;
- quality of the reproduced model (in comparison with the original);
- what problems (if any) were encountered.

*4) Communication Tests:* In the subjective communication test the user is asked how easy it is to communicate over the video part of the MAVT link. Is communication easy and flowing, as it would be if the two users were in the same room, or does the video link hinder normal communication? If communication is hindered, how is the hindrance experienced?

### B. Audio Subjective Tests

*1) Current Test Methods:* Several methods of subjective evaluation exist for evaluation of codec quality. These methods can be divided into three categories: articulation and diagnostic techniques, listener opinion tests, and conversation opinion tests. Articulation and diagnostic tests are very useful when the quality of the system is rather poor. Among these methods are the distinctness test [34], diagnostic rhyme test (DRT) [35], modified rhyme test (MRT) [36], and diagnostic acceptability measure (DAM) [37]. At least three different types of listener opinion tests exist. The absolute category rating (ACR) procedure [38] is used for evaluating medium-quality codecs. The degradation category rating (DCR) procedure [39] offers a high distinguishing level when used for high-quality coders. Comparison tests, like the ranked comparison test, are derived from the DCR test and also offer a high distinguishing level.

Conversation opinion tests evaluate the quality of the system when used in a more or less normal conversation. A large number of tests with varying degrees of artificiality have been designed. Probably the best known is the CCITT "picture sorting test" [40].

The main shortcoming of existing test methods is that usually equipment is tested instead of a complete communication link. Furthermore, these test methods assess the equipment from a technical viewpoint. It would be more logical to assess the equipment from the (subjective) point of view of the users. This idea has led to the development of a new set of tests, especially designed to assess a bidirectional communication link.

*2) A Comprehensive Test Approach:* The main objective is the design of a test set that would assess any degradation on a communication link. This set should be complete in the sense that any possible degradation is assessed, and in the sense that no more tests should be done than are absolutely necessary. The question that was constantly posed during the development of the test set was: "Which degradation will hinder either talker or listener or both in communication?"

The proposed test set is called the ultimate test set (UTS) [41] and is divided into five subsets. From each of these subsets, the communication link is assessed from a different point of view.

1) The objective test comprises the usual measurements that can be made on telephony equipment plus a few tests that are specific for an end-to-end bidirectional communication link.
2) The talker test comprises tests that should reveal any degradations that will hinder the talker in speaking. Degradations assessed are: talker echo, talker sidetone, and line disturbances.
3) The listener test aims to reveal any degradations that will hinder the listener in listening. Degradations assessed are: listener echo, listener sidetone, line disturbances, voice (center) clipping, listening effort, and the overall subjective quality.
4) The interaction test is designed to assess the communication link when talker and listener are involved in a highly interactive task, where the respective roles of talker and listener are frequently swapping. Degradations assessed are: voice clipping, delay, and voice suppression.
5) The conversation test should reveal the subjective quality of the link when it is used in a normal conversation. As the aspects assessed in the previous three test subsets are also key ingredients of this test, it is expected that the conversation test will offer a confirmation of the previous test results. Degradations assessed are: communication effort, naturalness of speech, the need to speak carefully, and the overall communication quality.

## IV. Procedures for the Subjective Tests

The field tests that have been conducted to date are based on those documented in the reference document [42]. Several alterations were made to enable the tests to take place on a single MAVT terminal. The tests are based on the system being in "loop-back" mode.

### *A. Experimental Environment Conditions*

The demonstrator PC was placed on a standard-height desk with the monitor positioned above the system unit. The camera was placed to the right-hand side of the monitor at the same level as the center of the monitor screen. The subject sat in front of the demonstrator slightly to the right-hand side. This gave the subject some desk space in front of them to move the mouse, fill in the questionnaire, and perform the tests. The monitor was angled slightly to the right to present the subject with a fully facing image. The distance between the subject, while seated, and the camera/monitor was about 1 m. During the tests the subject was allowed to sit in whatever position was comfortable.

Behind the subject, about 3 m from the monitor, there was a fixed blue screen which filled the entire camera view. Lighting was provided by four standard-length 80-W fluorescent tubes on the ceiling (ceiling height about 3 m). The microphone was positioned on the desk in front of the demonstrator. The loudspeaker was positioned to the right-hand side of the subject at a distance of about 50 cm. This distance between the loudspeaker and microphone was kept artificially large. This strategic position was used to avoid positive audio feedback that otherwise could occur.

The background noise present was typical of that found in a quiet office and should not affect the subject's performance. There were the sounds from several personal computer cooling fans, doors opening and closing in the background, the faint murmur of voices from nearby rooms, and negligible noise from outside the office. Other people present in the laboratory were asked not to make any loud sudden noises while the audio tests were being carried out. This was because the audio sequences were only played to the subjects once and it would not be possible to playback sections of the test sequences if they were obscured by loud noises.

*1) DECT Environment:* For the duration of the tests, the quality of the DECT radio channel was considered ideal (BER $<= 10^{-6}$). The test sequences have all been recorded over a clean error-free channel with similar quiet office background noise present. In the first subjective test "First Impressions," a live radio link with possible channel errors was used. The radio link had the transmitter and receiver in the same room at about 2-m separation.

*2) ATDMA Environment:* For the limited ATDMA tests it was not possible to achieve a real-time connection to the demonstrator built within the ATDMA project. In place of this, error patterns taken from the project were made available. These error patterns were overlaid over the clean DECT channel. This gave a simulated UMTS channel which could be used in the same way as the DECT subjective tests. The simulated UMTS channel had errors equivalent to a 15-dB SNR channel.

### *B. Quality Assessment Methods Used*

The reference paper [43] examines three different methods for the comparison of subjective test methods. The absolute category rating (ACR) method is generally used when differences between the reference source and the codec output are large. In the ACR method, a subject gives an opinion on a presentation of a stimulus without having access to a reference. The main drawback for the ACR method is that the subject will use his common knowledge of the world for creating his own ideal reference. The two other methods shown in the paper are the degradation category rating (DCR) in which the subject gives an opinion using a known reference stimulus and the two alternative forced choice method (2AFC) in which the subject gives an opinion of one of two choices. The DCR and 2AFC methods are usually used when small differences in quality have to be assessed. As the MAVT employs low bit rate audio and video coding algorithms, it is clear that there will be large impairments in the coded outputs. This leads to the conclusion that the ACR method is the best choice for the subjective testing of the system. Results from

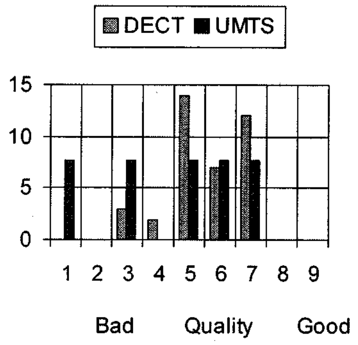


Fig. 5. Quality of the connection for DECT and UMTS.

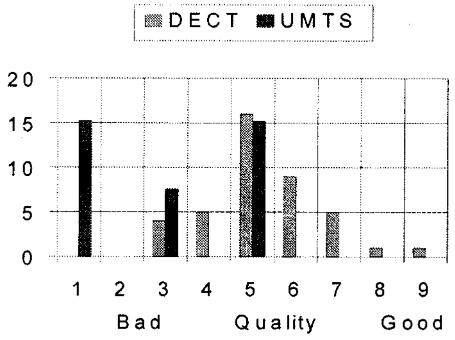


Fig. 6. Quality of the video for DECT and UMTS.

the paper show that if the number of subjects is large (31 subjects), then the results of all three methods are equally reliable.

### C. The Subjective Tests

There were two questionnaires involved in the subjective testing. The first was given to the test subject. This contained instructions on what the subject was expected to do and provided space for the subject to insert their answers. The second was for the use of the operator. This contained the correct answers to the tests.

*1) First Impressions:* For this first test, the subjects were given the opportunity to play with the MAVT for a few minutes. The operator initiated an MAVT connection and the subject was asked to sit in front of the MAVT terminal. The operator then explained what had happened and what was currently being shown on the monitor. The subject was then shown how some of the basic MAVT system functions work. They were then left to their own devices for a few minutes after being asked to pay special attention to the quality of the audio, video, and overall system characteristics.

At the end of this period, the operator returned and requested that the user complete the first group of questions on the questionnaire. These questions are answered by ticking a particular box on a nine point ACR scale.

i) *How would you rate the* **quality of the connection?** (Fig. 5.)

ii) *How would you rate the* **quality of the video** *part of the connection?* (Fig. 6.)

iii) *How would you rate the* **quality of the audio** *part of the connection?* (Fig. 7.)

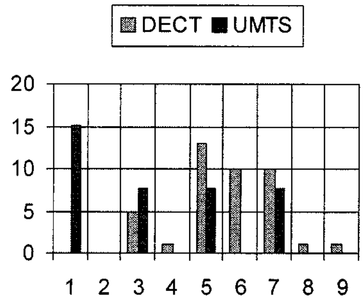


Fig. 7. Quality of the audio for DECT and UMTS.

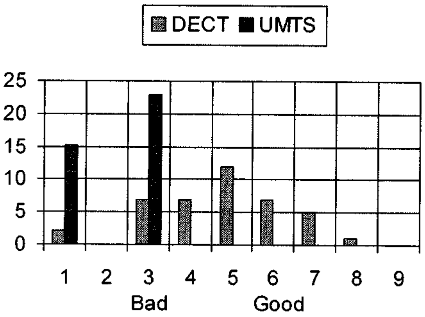


Fig. 8. Quality of the synchronization for DECT and UMTS.

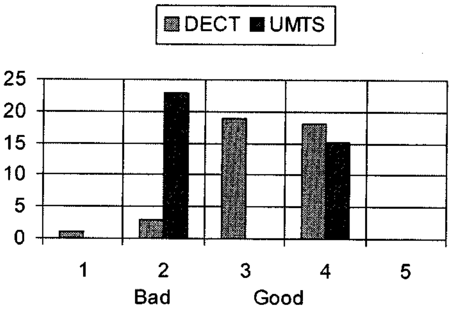


Fig. 9. Perception of the delay for DECT and UMTS.

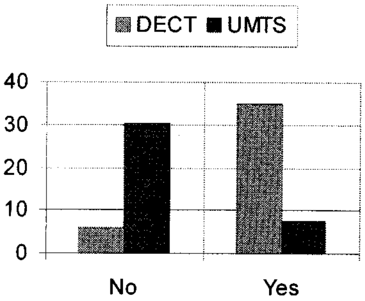


Fig. 10. Acceptability of the videophone system for DECT and UMTS.

iv) *How would you rate the* **synchronization** *between audio and video?* (Fig. 8.)

v) *How would you rate the* **delay?** (Fig. 9.)

vi) *Assuming that it was used for a mobile video telephone would you consider that it was acceptable?* (Fig. 10.)

*2) Object Identification and Placement:* In this test, the subject was shown a prerecorded video sequence in which various objects were placed one at a time on a flat surface. The subjects

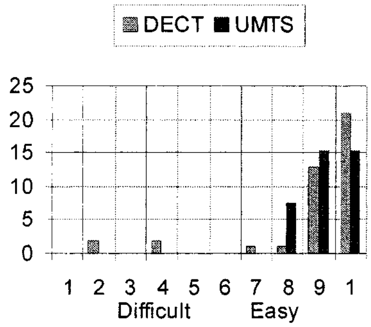


Fig. 11. Ease of object identification for DECT and UMTS.

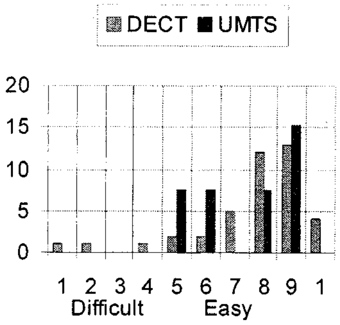


Fig. 12. Ease of story comparison for DECT and UMTS.

were given a selection of objects which included all the objects used in the video sequence and several other similar objects. It was the subject's task to duplicate the placement of the objects, as shown in the video sequence, on the table surface in front of them. The subjects were reminded to choose the correct object and to place it in the correct location relative to all the other objects used. They were given the time taken for the sequence to run to complete the task. The test was marked by giving one mark for each correctly used object and one mark for correctly positioning that object. The performance results are shown in Fig. 11. When the test was completed, the subjects were asked to indicate on an ACR scale of one–ten how easy they found the test to complete.

vii) *How easy was it to duplicate the pattern of objects?* (Fig. 15.)

The test objects are stored in a box for convenience. The objects are as given at the bottom of the page.

*3) Listen and Answer:* In this test, the subject was asked to listen to an audio sequence. There was a video sequence shown along with the audio sequence, but the subjects were asked to concentrate on the audio sequence only. The subjects were given six questions in the questionnaire which must be answered to the best of their ability after, or while, the audio sequence was played. The audio sequence consisted of a spoken story which lasted about 30 s. The text was spoken in what can be described as a normal speaking voice by a native speaker. The performance results are shown in Fig. 14. When the test was completed the subject was asked to indicate on an ACR scale of one–ten how easy they found the test to complete.

viii) *How easy was it to understand the spoken text over the audio link?* (Fig. 18.)

*4) Story Comparison:* This test was another audio-only test. The subject was given an audio sequence to listen to. There was a corresponding video sequence, but the subject was asked not to pay attention to it. The subject was given a near copy of the spoken text in the questionnaire. This passage was similar to the spoken text except that a few of the words have been changed. The subject has two tasks to perform. First, they should indicate which of the spoken words has been changed. Second, they should indicate what new words are used in place of the old ones. The subjects were not told how many words were changed. The performance results are shown in Fig. 12. When the test was completed, the subject was asked to indicate on an ACR scale of one–ten how easy they found the test to complete.

ix) *How easy was it to understand the spoken text over the audio link?* (Fig. 16.)

*5) Model Building:* This was the final sequence shown to the subject. In this test, the subject was asked to build a model out of LEGO bricks. They were given the LEGO bricks to examine before the start of the test. The subject was then shown an audio and video sequence. The sequence instructed the subject on how to construct the model using the given LEGO bricks. The subject was reminded to use both the audio and the video information to help in the construction of the model. At the end of the sequence, the subject was marked according to how many of the bricks were correctly placed. The performance results are shown in Fig. 17. When the test was completed, the subject was

| Correct object | Similar objects also in the box |
|---|---|
| Blue felt-tip pen | Black felt-tip pen |
| Small reel of tape | Large reel of tape |
| Black ball point pen | Blue ball point pen |
| Padlock | Security screw lock |
| Light green highlighter pen | Dark green highlighter pen |
| Grey pencil | Long screwdriver |
| 3.25 inch diskette | |
| Oval piece of paper inscribed “text2." | Oval piece of paper inscribed “text" |
| Oval piece of paper inscribed “hello." | Oval piece of paper inscribed “large." |

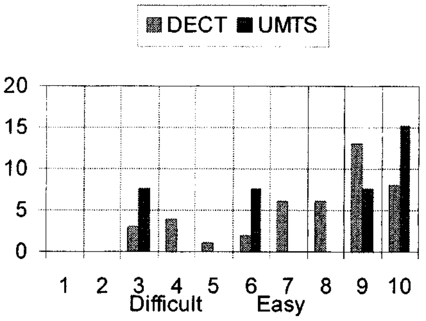


Fig. 13. Ease of model building for DECT and UMTS.

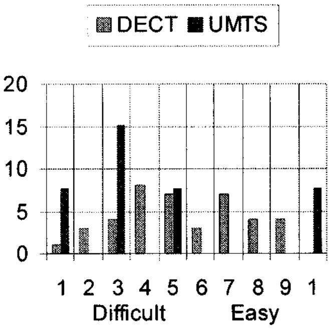


Fig. 14. Ease of listen and answer for DECT and UMTS.

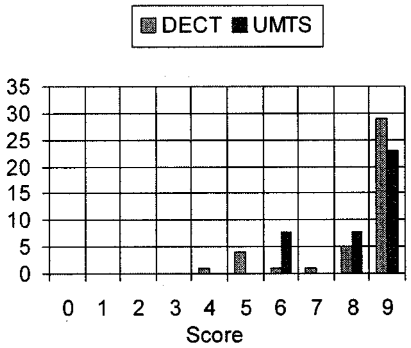


Fig. 15. Object identification performance for DECT and UMTS.

asked to indicate on an ACR scale of one–ten how easy they found the test to complete.

x) *How easy was it to build the model?* (Fig. 13.)

The model, a Reindeer, is made up of 17 LEGO pieces:

- 6 white blocks with dimensions: 8 × 9.5 × 16 mm;
- 2 white blocks with dimensions: 9.5 × 16 × 32 mm;
- 1 white block with dimensions: 3 × 16 × 32 mm;
- 4 white wedge-like blocks with dimensions: block—8 × 9.5 × 8 mm + wedge—8 × 9.5 × 8 mm;
- 1 white wedge-like block with dimensions: block—8 × 9.5 × 16 mm + wedge—8 × 9.5 × 16 mm;
- 1 white wedge-like block with dimensions: block—8 × 9.5 × 16 mm + wedge—16 × 9.5 × 16 mm;
- black quarter-circle pieces with dimensions: 8 × 9.5 × 8 (inner radius) and 16 (outer radius) mm.

*6) Exit Questionnaire:* After the tests with the demonstrator finished, the subject was asked to fill out an exit questionnaire. This allows the subject to make known any points which they feel are important. The subject was requested to fill in as much detail as they thought necessary.

xi) *Which factors would you consider sufficiently important that you would like to see improved?*
xii) *Which were the most annoying impairments in the video part?*
xiii) *Which were the most annoying impairments in the audio part?*
xiv) *Which were the most annoying impairments of the system in general?*
xv) *How user friendly was the user interface?*
xvi) *Do you have any other comments you wish known to us?*

## V. Analysis of Results in DECT and UMTS

For DECT, the group comprised of 42 test subjects was taken from the general university population. A population of 42 was accepted as statistically significant [43].

Unfortunately, it was not possible to perform real-time field tests using the UMTS hardware test bed from the ATDMA project. It was decided that the best way to proceed was to perform our own trials using simulated error patterns generated by the ATDMA project. The error patterns that were generated were for a 15-dB SNR. To generate our simulated UMTS channel the same data sequences used by the DECT field tests were rerecorded using the ATDMA error patterns which were overlaid on top of a clean outgoing DECT channel. In this way, the only errors recorded on the data sequences were those caused by the error patterns. When these were played back directly onto the decoders only UMTS errors were seen and heard. The channel codec for the audio and video codecs was adjusted to be optimized for the new channel conditions by measuring the statistics of the error patterns (mean and autocorrelation of the error patterns) and using these statistics to parameterize the channel codecs.

Due to time constraints it was also not possible to perform the UMTS simulations with a large number of naive users in the same way as the DECT trials had been performed. It was also seen during the experimental setup that the results for the errors on the UMTS system caused quite a degradation in the audio and video quality. After all the considerations and constraints had been noted, it was decided to perform a limited set of field tests using expert users. An expert user was defined as a person who was experienced in viewing/listening to low bit rate video/audio and who understood the associated problems. The users gauged the effect of errors on the system and gave an impression of how well the system behaved. In this way, it was hoped that some insight into how the terminal behaved in a UMTS environment would be gained. If the tests were carried out using naive users, then it was possible that they would have a low tolerance to errors and would give bad scores. These scores would just say that the system was not adequate for public use. What they would not give is an impression of how well the system worked considering the fact that it was of low bit rate and operated over an error prone channel. The simulated channel had an SNR of 15 dB which corresponded to an error rate of about $3.16 * 10^{-2}$.

### A. Quality of the Connection

In general, the DECT naive users gave favorable scores to the MAVT with averages at the median of the score ranges for quality of connection (Fig. 5), quality of the video (Fig. 6), quality of audio (Fig. 7), and perception of delay (Figs. 8 and 9). However, there was always a spread of scores throughout the whole range for all tests. Despite this, the majority of the DECT naive users found that the MAVT was acceptable to use.

In contrast, the UMTS expert users gave less favorable scores to the MAVT with averages toward the lower end of the score ranges for quality of connection, quality of the video, quality of audio, and perception of delay. Correspondingly, the majority of the UMTS expert users found that the MAVT (Fig. 10) was unacceptable to use.

### B. Subjective Tests: Ease of Use

These results showed how easy the test subjects thought the different subject tests were to perform. They were asked to give their opinion without knowing how they had scored in each test. Therefore, the results in this section were based purely on how well the subjects thought they had performed and not on how they actually performed.

In general, the DECT naive users gave very favorable scores for their perception of how they performed in the tests with averages at the higher end of the score ranges for ease of object identification (Fig. 11), ease of story comparison (Fig. 12), and ease of model building (Fig. 13). However, their perception of how they performed for ease of listen and answer test (Fig. 14) were less favorable with a score average at the median of the range. This indicates that the users perceived the system easy to use to carry out the different subject tests for those tests that did not need to use comprehension of speech. The UMTS expert users perception of how they performed in the tests were very similar to those of the DECT naive users.

### C. Test Performance

These results show how the test subjects actually performed for each of the different tests.

The DECT naive users actual and perceived performances were highly correlated. They obtained very favorable scores with averages at the higher end of the score ranges for ease of object identification (Fig. 15), ease of story comparison (Fig. 16), and ease of model building (Fig. 17). They obtained less favorable scores with averages at the median of the score ranges for ease of listen and answer test (Fig. 18). Similarly, the UMTS expert users actual and perceived performances in the tests were very similar to those of the DECT naive users.

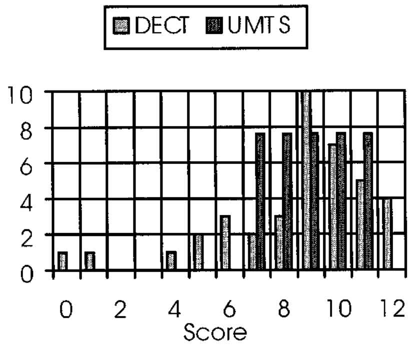


Fig. 16. Story comparison performance for DECT and UMTS.

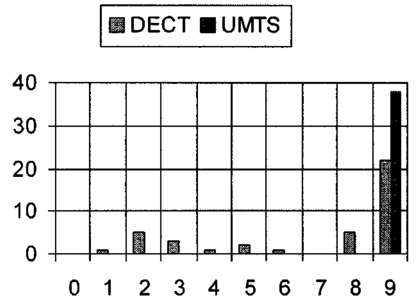


Fig. 17. Model building performance for DECT and UMTS.

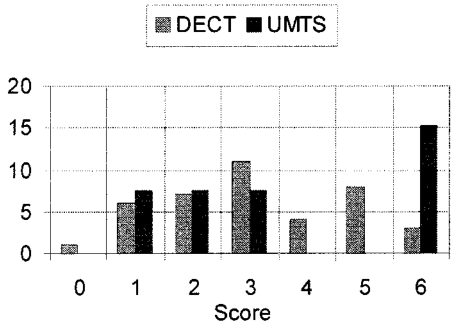


Fig. 18. Listen and answer performance for DECT and UMTS.

## VI. Conclusions of the Subjective Tests

When using a DECT network, the MAVT produced reasonable audio and video quality despite the poor quality of connection, long delays, and poor synchronization between speech and video. General comments from the test subjects all suggested that the video and audio quality needed to be improved despite providing an acceptable video quality as indicated by the subjective tests. The results from the test sequences showed that it was possible to successfully use audio/video terminals to communicate with each other over a 32-kb/s DECT network link. Therefore, even if the quality was not good enough for general use, the system could still be used in situations where high-quality audio and video was not necessary. It was shown that the quality of the audio was sufficiently good for comprehension and the video was sufficiently good for following and repeating simple mechanical tasks. However, the quality of the video was not sufficiently high for general use. Therefore, a link capacity that was greater than 32 kb/s would be required from the network for an improvement of the subjective video quality.

The test results in a simulated UMTS network suggested that the MAVT was not acceptable as a video telephone system in its present form. More effort would be required to improve the audio and video quality with respect to its ability to handle mobile radio channels with high-error rates. Even though the system scored rather badly in terms of quality, the results of the test sequences demonstrated that useful information could still be communicated over the erroneous radio network. Therefore, the terminal may still be able to find its place in certain professional applications where conveying poor quality video information at low bit rates is of importance.

In order to improve the quality of the video, a number of further measures were suggested. The most obvious improvement could be obtained by improving the coding efficiency. However, it is clear that improving coding efficiency on its own may not provide sufficient compression for good quality video. Content-based scalability combined with object coding of images will provide the ability to achieve scalability with a fine granularity in spatial resolution, temporal resolution, quality, and complexity for the different objects. Content-based bit stream multiplexing will allow selected objects within the image to be transmitted when there is insufficient capacity in the transmission link.

A further serious limitation of the MAVT system is that it is very inflexible. The audio, video, and error control algorithms are fixed once they have been embedded onto the hardware. There is no means of modifying the functionality of the system once it is up and running. This means, for example, that error coding algorithms are always parameterized according to the worst channel characteristic that is expected. Future terminals may be able to adapt themselves by using an adaptive syntax. For example, if a called terminal did not have a particular function it would request that function from the calling terminal. After receiving the new function, the called terminal would reconfigure itself such that this new function could be used. This will enable terminals to be adaptable, compatible, and expandable.

The MAVT is the first system to demonstrate real-time moving video and audio over a low bit rate mobile radio channel. The subjective tests have shown that, although the audio and video quality are not of a very high standard, the quality is more than adequate for most intelligible communication requirements.

## REFERENCES

[1] N. MacDonald, "Transmission of compressed video over radio links," *BT Technol. J.*, vol. 11, no. 2, pp. 182–185, 1993.

[2] ——, "Transmission of compressed video over radio links," *SPIE Visual Commun. Image Processing*, vol. 1818, no. 3, pp. 1484–1488, 1992.

[3] R. Schafer, "European R & D activities in mobile multimedia communications and expectations on future services," *IEICE Trans. Commun.*, vol. E77-B, no. 9, pp. 1083–1088, 1994.

[4] J. Cosmas, B. Evans, C. Evci, W. Herzig, H. Persson, J. Pettifor, P. Polese, R. Rheinschmitt, and A. Samukic, "Overview of the mobile communications program of RACE II," *Electron. Commun. Eng. J.*, vol. 7, no. 4, pp. 155–167, 1995.

[5] D. A. Gill, "The MAVT in a mobile radio network environment," in *RACE Mobile Telecommunication Workshop*, 1993, pp. 198–201.

[6] J. P. Cosmas, A. J. Pearmain, D. Gill, and J. Zouain, "Mobile audio–visual terminal for DECT mobile radio system," in *5th Bangor Symp. Telecommunications*, 1993, pp. 187–190.

[7] P. Crespo, J. Cosmas, N. Condette, and R. Mann-Pelz, "Channel error profile for DECT," RACE Document R2072/TEL/1.2/DS/R/001, 1992.

[8] "ATDMA RTTB-MAVT terminal interconnection," RACE Document R2084/SM/FEL/DN/R/001/a1.

[9] "Half burst mode for transport of 32 Kbit/sec," RACE Document R2084/AMCF/TI3/IN/I/240/a1, 1995.

[10] "Proposal for RTTB implementation work plan," RACE Document R2084/NOK/TI5/IN/I/108/a1, 1995.

[11] D. Lappe, "Flexible video codec," RACE Document R2072/BOS/2.1/DS/R/016, 1993.

[12] G. Nitsche, "Implemented algorithm for low bit rate (DECT $p = 2$)," RACE Document R2072/BOS/3.1/DS/I/023/b1, 1994.

[13] "Video codec for audio–visual services at px64kbit/s," CCITT Recommendation H.261, 1989.

[14] "Video coding for narrow telecommunication channels at <64 kbit/s," ITU-T SG15 Draft Recommendation H.26p, 1994.

[15] R. G. White, "Compressing image data with quadtrees," *Dr. Dobb's J Software Tools*, vol. 12, no. 3, pp. 16–45, 1987.

[16] G. G. Langdon, "An introduction to arithmetic coding," *IBM J. Res. Develop.*, vol. 29, no. 2, pp. 135–149, 1984.

[17] A. N. Netravali, *Digital Pictures: Representation and Compression*. New York: Plenum, 1988.

[18] C. Stiller, "Motion-estimation for coding of moving video at 8kbit/s with Gibbs modeled vector field smoothing," in *SPIE Visual Communications and Image Processing*, Lausanne, Switzerland, 1990.

[19] A. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Commun. ACM*, vol. 30, no. 6, pp. 520–540, June 1982.

[20] "Description of reference model 8," COST211BIS/SIM89/37, 1989.

[21] "Reference model 8 (RM8)," CCITT SG15, 1989.

[22] J. Hagenauer, "Rate-compatible punctured convolutional codes (RCPC codes) and their applications," *IEEE Trans. Commun.*, vol. 36, no. 4, pp. 389–400, 1988.

[23] R. Mann-Pelz, "An unequal error protected px8 kbit/s video transmission for DECT," in *IEEE Proc. Veh. Technol. Conf., VTC'94*, Stockholm, Sweden, Sept. 1994.

[24] P. Crespo, J. Garcia-Frias, R. Mann-Pelz, and P. Mege, "UMTS channel coding for the MAVT," RACE Document R2072/TEL/7.2/DR/L/041/a, 1995.

[25] S. Mayer and J. Boudy, "Speech processing algorithms," RACE Document R2072/MATRA/2.2/DS/S/008/b1, 1993.

[26] S. Scott, "Speech coding software description," RACE Document R2072/MATRA/WP3.2/DS/C/029/b1, 1993.

[27] M. M. Sondhi, "An adaptive echo canceller," *Bell Syst. Tech. J.*, vol. 46, no. 3, pp. 497–511, Mar. 1967.

[28] J. S. Lim and A. Oppenheim, "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proc. IEEE*, vol. 67, no. 12, 1979.

[29] D. B. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, pp. 786–794, Aug. 1981.

[30] M. Delprat, M. Levet, and C. Gruet, "A 6kbps regular pulse CELP for mobile communications," in *Advances in Speech Coding*. Norwell, MA: Kluwer Academic, 1991.

[31] P. Vary, "Noise supression by spectral magnitude estimation mechanism and theoretical limits," *Signal Processing*, vol. 8, pp. 387–400, 1985.

[32] D. B. Boll, "Supression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, Apr. 1979.

[33] M. Roux, R. Mann Pelz, and P. Crespo, "Error correction for video and speech coding (px8 kbits)," RACE Document R2072/BOS/2.1.3/DR/L/014/b1, 1993.

[34] CCITT, "Mesure de l'affaiblissement equivalent pour la nettete d'un systeme telephonique commercial par comparaison avec ls systeme de reference SRAEN," *Avis p. 43*, vol. V, pp. 69–114, 1960.

[35] W. D. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technol.*, vol. 1, no. 4, pp. 30–39, Jan./Feb. 1983.

[36] A. S. House *et al.*, "Articulation testing methods: Consonantal differentiation with a close-response set," *J. Acoust. Soc. Amer.*, vol. 37, no. 1, pp. 158–166, Jan. 1965.

[37] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[38] CCITT, "Absolute category rating (ACR) method for subjective testing of digital processes," in *Blue Book*, ser. P recommendations. Geneva, Switzerland: ITU, 1989, vol. V, pp. 346–351.

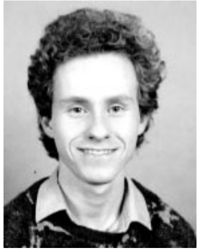
[39] ——, "Subjective performance assessment of digital encoders using the degradation category rating procedure (DCR)," in *Blue Book*, ser. P recommendations. Geneva, Switzerland: ITU, 1989, vol. V, pp. 351–356.

[40] ——, "Methods used for assessing telephony transmission performance," in *Blue Book*, ser. P recommendations. Geneva, Switzerland: ITU, 1989, vol. V, pp. 237–246.

[41] G. M. Loose and P. T. Pont, "The ultimate test set," contribution to RACE R2072, MAVT, 72/PTT Research/WP2.2/DN/C/22.11.93/2.1, Nov. 1993.

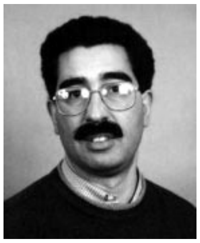
[42] A. Carvalho, A. Pearmain, W. Vogt, F. Mundt, and L. Contin, "Definition of DECT/UMTS field tests," MAVT Deliverable R2072/QMWC/WP4.2/039/#7.003, Dec. 1994.

[43] M. van Dort, J. G. Beerends, W. van den Brink, M. Loose, and L. Contin, "Comparison of three subjective video quality assessment methods," in *RACE Mobile Telecommunications Summit*, Cascais, Portugal, Nov. 1995.



**David Gill** received the B.S. degree in electronic engineering (second class honors) in 1991 and the Ph.D. degree in 1998, both from Queen Mary and Westfield College, University of London, London, U.K.

After graduating, he was a Research Assistant in the Telecommunications Research Department, University of London. During this time, he worked on the Research into Advanced Communications in Europe (RACE) project MAVT. He is currently with the Communication Systems Division, ERA Technology Limited, Leatherhead Surrey, U.K.



**John Paul Cosmas** (M'90) received the B.Sc.(Eng.) degree with honors in electronic engineering from Liverpool University, Liverpool, U.K., in 1978 and the Ph.D. degree in image processing and pattern recognition from Imperial College, U.K., in 1987.

From 1978 to 1983, he was an Electronics Development Engineer at Tube Investments and Fairchild Camera and Instruments. In 1983, he joined Imperial College as a Research Student and in 1986 was a Lecturer in digital systems design and telecommunications at Queen Mary and Westfield College, University of London, London, U.K. He is currently with the Department of Electronic and Computer Engineering, Brunel University, Middlesex, U.K. He has contributed toward the EEC research programs R1022 Technology for ATD, R2072 Mobile Audio–Visual Terminal, AC098 Mobile Multimedia Systems, and AC30073 CustomTV as well as the U.K. research project ATM Resource Management. His research is concerned with digital image processing and multimedia systems for telecommunication systems.

**Alan Pearmain** (M'80) received the B.Sc.(Eng.) degree in electrical engineering in 1967 and the Ph.D. degree in topic liquid-insulated electrostatic generators in 1971, both from Southampton University, Southampton, U.K.

He was a Research Fellow at Heriot-Watt University, Edinburgh, U.K., from 1970 to 1972 and a Lecturer at University College, Dublin, U.K., from 1974 to 1979. He was on a one-year sabbatical from the University College and went to the Brookhaven National Laboratory, NY, from 1977 to 1978, where he worked on the superconducting power transmission project. He has been a Member of the Electronic Engineering Faculty, Queen Mary and Westfield College, University of London, London, U.K., since 1979. During this time, he has pursued research in areas of high voltage, dielectric liquids, VLSI CAD tools, chip architectures, ATM test equipment, and mobile multimedia terminals. He is currently working on the European Union research projects mobile multimedia systems and CustomTV.