Notiziario Tecnico

Archivio

3/1997

TIM

NOTIZIARIO TECNICO TELECOM ITALIA

Notiziario Tecnico Telecom Italia
Anno 6 - n. 3 - Dicembre 1997

Editore
Telecom Italia S.p.A.

Direttore responsabile
Rocco Casale

Comitato di direzione
Luigi Bonavoglia, Claudio Carrelli, Maurizio Dècina, Umberto de Julio, Giuseppe Gerarduzzi, Cesare Mossotto, Giorgio Pellegrini, Aldo Roveri, Paolo Rumboldt, Carlo Giacomo Someda, Francesco Tisato, Francesco Valdoni
Segretario: Stefano Mariani

Comitato tecnico
Claudio Boreggi, Claudio Brosco, Odoardo Brugia, Emilio Cancer, Giuseppe Carra, Gianfranco Ciccarella, Roberto Colantonio, Duccio Di Pino, Giuseppe Grimaldi, Giancarlo Miranda, Paolo Oberto, Romolo Pietroiusti, Lorenzo Roberti Vittory, Diego Zandel
Segretario: Stefano Mariani

Segreteria tecnica
Andrea Baiocchi

Redazione
Coordinamento: Romolo Pietroiusti; Claudio Brosco, Flavio Cataldi, Enzo Garetti, Piero Izzo, Giuseppe Giacobbo Scavo

Segreteria di redazione
Francesca Romana Belgiovine, Daniela Ceccarelli

Progetto grafico e copertina
Modo di Luca Modugno

Grafica ed impaginazione
Maurizio Feliciangeli, Modo, FC Studio

Fotografie
Musei Vaticani (P. Zigrossi) e gentilmente fornite dal Professor Andrew J. Viterbi, dalla Ericsson e dall'Istituto Nazionale per la Grafica

Consulenza Redazionale
Claudia Bonamano, Adriano Santelli, Cinzia Vetrano

Stampa
Union Printing - Viterbo

Registrazione
Periodico iscritto al n. 00322/92 del Registro della Stampa presso il Tribunale di Roma in data 20/05/92

Direzione e Redazione
via di Val Cannuta, 250 - 00166 Roma
tel. +39+6+3688-3801 - fax +39+6+6633035

Andrew J. Viterbi riceve il diploma di laurea Honoris Causa.

SOMMARIO

Ai Lettori

Fermiamoci un momento per riflettere

Nel corso di un'escursione impegnativa in montagna, quando la stanchezza fisica comincia a farsi sentire, ci piace fermarci, voltarci indietro e osservare il cammino percorso; guardiamo poi verso l'alto per cercare di scoprire le difficoltà che ci aspettano ancora e per scorgere, se possibile, la vetta. La breve pausa per prender fiato ci permette così di proseguire con maggior lena e con maggior convinzione la salita.

Proviamo le stesse sensazioni se ci fermiamo a guardare il cammino compiuto dalle telecomunicazioni in questo secolo e, in particolare, se ripercorriamo la rapida evoluzione che si è avuta in questi ultimi anni e se ci soffermiamo sugli uomini che sono stati protagonisti dei cambiamenti; i successi del passato ci spingono poi a guardare in avanti, a cercare di "leggere il futuro"; a tentare cioè di individuare gli scenari che si delineano nel breve e nel medio termine per il settore d'avanguardia nel quale oggi operiamo.

Ci siamo fermati per qualche ora il 13 maggio di quest'anno, quando l'Università degli Studi Tor Vergata di Roma e l'AIIT (l'Associazione Italiana degli Ingegneri delle Telecomunicazioni) hanno organizzato nello stesso giorno due manifestazioni di assoluto rilievo: il conferimento della laurea Honoris Causa ad Andrew J. Viterbi e un seminario sulle "Telecomunicazioni verso l'assetto numerico".

Due brevi note a commento di questa giornata. La laurea Honoris Causa al Professor Viterbi segue di pochi mesi quella che l'Università di Padova ha conferito al Professor Charles K. Kao il 18 ottobre dello scorso anno: Kao per la sua attività di ricerca legata alle fibre ottiche; Viterbi per quella connessa alle trasmissioni radio. Con questi importanti riconoscimenti l'Università italiana ha voluto sottolineare il ruolo di due protagonisti della ricerca che hanno dato un particolare impulso all'innovazione della tecnologia dei portanti trasmissivi oggi impiegati per il trasferimento dell'informazione a distanza.

Senza indulgere alla ricerca di una matrice nazionale per i brillanti risultati di studi, dato che Viterbi dovette lasciare l'Italia in giovanissima età, ci piace credere che questo scienziato abbia portato con sé in Canada e negli Stati Uniti, ove ha svolto prevalentemente la sua attività di studioso, un po' della nostra cultura; e che questa abbia contribuito - in misura forse non grande ma pur sempre significativa - ai suoi successi e ai numerosi riconoscimenti che gli sono stati conferiti da prestigiose istituzioni di molti Paesi. L'ultimo in ordine di tempo è quello ricevuto in questa occasione.

Con un po' di stupore di alcuni tra i presenti alla cerimonia, e con l'emozione di tutti, la pur giovane Università di Tor Vergata ha voluto che il conferimento della laurea Honoris Causa si svolgesse in latino riprendendo, in un'occasione di indubbia rilevanza, anche simbolica, una tradizione che sottolinea il ruolo avuto per secoli dalla lingua latina, quale strumento di comunicazione universale tra i soggetti e gli istituti impegnati ai massimi livelli nello studio di tutte le principali discipline, e di espressione formale del pensiero umano.

Ai Lettori

Lo stesso Viterbi ha sottolineato prima della "Lectio Doctoralis" tenuta in questa occasione, di aver studiato da giovane in Canada il latino, ricordando come suo padre, che si era laureato in Italia, gli raccomandasse di non tralasciare gli studi classici.

Nel riportare nel Notiziario *il testo originale - in latino - del conferimento della laurea Honoris Causa, si è deciso di non ridurne la suggestione affiancando ad esso la traduzione in "volgare".*

Nel seminario sulle "Telecomunicazioni verso l'assetto numerico" è stato poi mostrato uno scenario approfondito e affidabile sulle prospettive delle evoluzioni dell'ICT (Information & Communication Technology) da esperti assai noti per le profonde conoscenze possedute e per le funzioni di alta responsabilità che svolgono nelle Società in cui operano. Si è quindi ritenuto opportuno proporre nel Notiziario *gli interventi dei tre relatori per permettere ai nostri lettori di conoscere gli orientamenti di grande interesse emersi in questa giornata.*

La rivista viene pubblicata, per motivi redazionali, alcuni mesi dopo la manifestazione. In uno scenario in continuo, rapido cambiamento, qualcosa è già mutata; ma si sono voluti conservare i testi degli interventi così come presentati, per fissare queste previsioni sul futuro a un particolare momento del cammino dell'ICT. L'auspicio è quindi che i lettori possano in qualche modo rivivere una giornata ricca di ricordi, di emozioni e di spunti di riflessione per il futuro.

Desidero da ultimo ringraziare, anche a nome del Comitato Direttivo della rivista, gli organizzatori della manifestazione e, in particolare, l'Università degli Studi Tor Vergata di Roma e l'AIIT per aver permesso e incoraggiato la pubblicazione sul Notiziario Tecnico Telecom Italia *dei testi riguardanti l'intera manifestazione.*

r.c.

Personaggi del corteo imperiale. (Ara Pacis di Augusto, Roma).

Ai lettori

Motivazione della Giornata

Francesco Valdoni, Professore di Comunicazioni Elettriche presso l'Università di Roma Tor Vergata, porge il saluto ai convenuti.

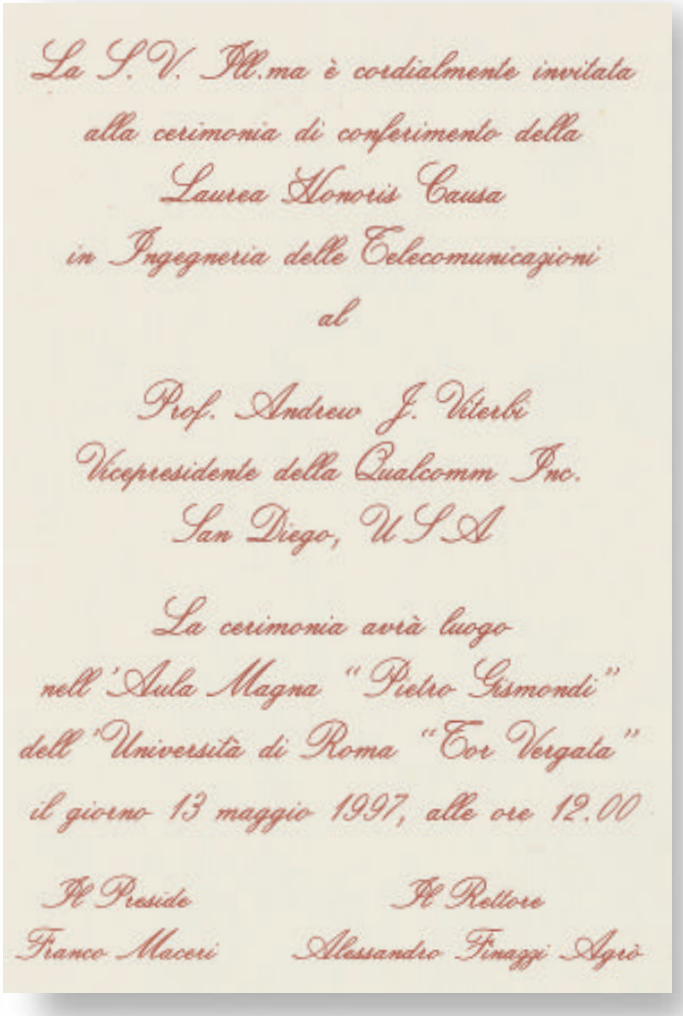
Andrew J. Viterbi nella sua lunga carriera ha fornito eccezionali contributi, sia in veste di scienziato che come imprenditore. Volendo concentrare l'attenzione in prevalenza sulla sua feconda attività di ricerca, si può anzitutto sottolineare che la produzione scientifica di Viterbi, straordinaria non solo per qualità avendo raggiunto risultati di assoluta eccellenza, è anche caratterizzata da una copiosa continuità, essendosi sviluppata senza interruzione su un arco di tempo di circa 35 anni. Grazie a questa opera, Viterbi occupa una posizione preminente nella storia dello sviluppo delle radiocomunicazioni in forma numerica.

Giovane ricercatore presso il Jet Propulsion Laboratory, in California, negli anni Sessanta Viterbi partecipò alla fantastica esperienza della conquista dello spazio interplanetario, che per la esigenza di radiocollegamenti tra sonde spaziali e la Terra a enorme distanza spinse alla ricerca di nuove tecniche di codifica dell'informazione trasmessa in forma numerica. Le formulazioni di Shannon alla base della teoria della informazione, apparse alla fine degli anni Quaranta, consentivano di prevedere esattamente l'ultima frontiera delle prestazioni nella trasmissione su canali rumorosi, mentre i progressi nella direzione di soluzioni tecniche effettivamente praticabili per avvicinare la menzionata frontiera rimasero modesti fino al 1961, quando Viterbi riuscì a stabilire, con il suo primo fondamentale contributo [1], come fosse possibile trasmettere senza errori una sequenza di cifre binarie, una volta adottato un codice ortogonale di lunghezza tendente all'infinito e superato un opportuno valore del rapporto segnale-rumore.

Nel seguito la ricerca di Viterbi si focalizzò sui codici convoluzionali, la cui potenzialità era sottostimata a favore di quelli a blocchi. Il risultato assai brillante di questa attività fu la introduzione, nel 1967, di un metodo di decodifica a traliccio [2], da allora universalmente noto come algoritmo di Viterbi, che consente di ottenere prestazioni assai prossime a quelle dei limiti teorici.

A partire dagli anni Settanta, Viterbi ha affiancato alle attività di ricerca e di didattica universitaria la partecipazione a iniziative imprenditoriali di successo, come cofondatore prima della Linkabit Corp., attiva nel segmento terreno di sistemi satellitari con piccoli terminali (VSAT, Very Small Aperture Terminal), e poi della Qualcomm Inc., società di telecomunicazioni mobili cellulari e via satellite. Il nuovo impegno non ha di certo impedito a Viterbi di continuare a fornire rilevanti contributi scientifici, sulle architetture modem

Ai lettori


La S.V. Ill.ma è cordialmente invitata
alla cerimonia di conferimento della
Laurea Honoris Causa
in Ingegneria delle Telecomunicazioni
al
Prof. Andrew J. Viterbi
Vicepresidente della Qualcomm Inc.
San Diego, USA
La cerimonia avrà luogo
nell'Aula Magna "Pietro Gismondi"
dell'Università di Roma "Tor Vergata"
il giorno 13 maggio 1997, alle ore 12.00
Il Preside
Franco Maceri
Il Rettore
Alessandro Finazzi Agrò

in tecnica numerica [4], sui metodi di sincronizzazione [5] e su numerosi aspetti, sia concettuali che applicativi, riguardanti i radiosistemi con accesso multiplo a divisione di codice (CDMA) [6] - [12].

Con l'amico Francesco Vatalaro e con altri più giovani colleghi del Dipartimento cui appartengo, abbiamo avuto la fortunata occasione di non pochi incontri con Andrew J. Viterbi, così da trasformare una prima conoscenza formale in un più approfondito rapporto culturale e, perché no, anche di gratificante contatto umano. Nonostante la vasta notorietà e il grande successo conseguiti, comprovati da numerosi riconoscimenti da parte di enti scientifici e culturali di grande prestigio internazionale, Viterbi ha conservato intatto il naturale atteggiamento, amichevolmente aperto verso ogni collega e serenamente autorevole, tipico dell'uomo di grande cultura che si adopera con entusiasmo per la diffusione del sapere. Ci è parso pertanto un dovere di proporre alla nostra Università di Roma Tor Vergata il conferimento a un tale personaggio della laurea Honoris Causa in Ingegneria delle Telecomunicazioni.

La cerimonia, che si è tenuta nell'Aula Magna dell'Ateneo lo scorso 13 maggio, ha offerto al Rettore Alessandro Finazzi Agrò e al Preside della Facoltà di Ingegneria Franco Maceri anche l'occasione (o forse il privilegio) di avere Andrew J. Viterbi quale primo laureato del menzionato Corso di Laurea in Ingegneria delle Telecomunicazioni, solo recentemente istituito presso Tor Vergata. La giornata da me introdotta è stata resa ancora più memorabile dallo svolgimento di un seminario sul tema Le telecomunicazioni verso l'assetto numerico*, organizzato dall'Associazione Italiana degli Ingegneri delle Telecomunicazioni e imperniato su brillanti relazioni svolte da Giovanni De Guzzis, da Umberto de Julio e da Guido Vannucchi.*

Francesco Valdoni

Ai lettori

ALCUNI TRA I PIÙ IMPORTANTI LAVORI DI ANDREW J. VITERBI

[1] Viterbi, A.J.: *On coded phase-coherent communications.* «I.R.E. Trans. on Space Electronics and Telemetry», Vol. SET-7, March 1961, pp. 3-14.

[2] Viterbi, A.J.: *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.* «IEEE Trans. Info. Theory», Vol. IT-13, 1967, pp. 260-269.

[3] Viterbi, A.J.: *Convolutional codes and their performance in communication systems.* «IEEE Trans. Com. Tech.», Vol. Com-19, 1971, pp. 751-772.

[4] Heegard, C.; Heller, J.A.; Viterbi, A.J.: *A microprocessor-based PSK modem for packet transmission over satellite channels.* «IEEE Trans. Commun.», Vol. COM-26, n. 5, May 1978, pp. 552-564.

[5] Viterbi, A.J.; Viterbi, A.M.: *Nonlinear estimation of PSK-modulated carrier phase with application to burst digital transmission.* «IEEE Trans. Info. Theory», Vol. IT-29, July 1983, pp. 543-551.

[6] Viterbi, A.J.: *When not to spread spectrum - A sequel.* «IEEE Commun. Mag.», Vol. 23, April 1985, pp. 12-17.

[7] Gilhousen, K.S.; Jacobs, I.M.; Padovani, R.; Viterbi, A.J.; Weaver, L.A. Jr.; C.E. Wheatley III: *On the capacity of a cellular CDMA system.* «IEEE Trans. Veh. Technol.», Vol. 40, n. 2, May 1991, pp. 303-312.

[8] Viterbi, A.J.: *Very low rate convolutional codes for maximum theoretical performance of spread-spectrum multiple-access channels.* «IEEE J. Sel. Areas Commun.», Vol. JSAC-8, May 1990, pp. 641-649.

[9] Viterbi, A.J.: *Wireless digital communication: a view based on three lessons learned.* «IEEE Commun. Mag.», September 1991, pp. 33-36.

[10] Viterbi, A.J.; Viterbi, A.M.; Zehavi, E.: *Performance of power-controlled wideband terrestrial digital communications.* «IEEE Trans. Commun.» Vol. 41, n. 4, April 1993, pp. 559-569.

[11] Viterbi, A.J.; Viterbi, A.M.; Zehavi, E.: *Other-cell interference in cellular power-controlled CDMA.* «IEEE Trans. Commun.», Vol. 42, n. 2/3/4, February/March/April 1994, pp. 1501-1504.

[12] Viterbi, A.J.; Viterbi, A.M.; Gilhousen, K.S.; Zehavi, E.: *Soft handoff extends CDMA cell coverage and increases reverse link capacity.* «IEEE J. Sel. Areas Commun.», Vol. 12, n. 8, October 1994, pp. 1281-1287.

Conferimento laurea Honoris Causa

Apertura della cerimonia e saluto di benvenuto

Universitatis Rector: Clarissimi Professores, Sodales universi!

Dies festus hic est, quo laetantes celebramus virum doctissimum, optime meritum de arte eminus electrica vi nuntiandi, Andream Viterbi dico, cui libenti animo salutem nuntio.

Alessandro Finazzi Agrò (a sinistra), Rettore presso l'Università di Tor Vergata di Roma assieme ad Andrew J. Viterbi.

Huius ritus vices hae erunt: primum Franciscus Maceri, Facultatis Decanus, de arte eminus nuntiandi vobis adloquetur, deinde Aldus Roveri, professor, laudationem magistri nostri aget, mox Andreas Viterbi ipse dissertationem suam dicet, ego demum, audita vestra sententia, diploma sollemne tradam recenti doctori.

Facultatis Decanus: Magnifice Rector, Professores Clarissimi, hospites Illustrissimi, alumni nobis dilecti, salvete!

Hoc die Studium nostrum artem aevi nostri recentem, eminus vi electrica nuntiandi dico, celebrat.

Rerum cognitio enim magna civilis cultus pars est putanda atque hominibus quam celerrime notitias et notiones conferre ipsum auget cultum.

Temporibus antiquis, Caesar ipse narratur tabellas in Foro exposuisse ut populo romano diurna eventa celeriter patefaceret.

Nonnulli, praeterea, rerum gestarum scriptores priscum romanum imperium cecidisse putant hac de causa inter caeteras, res in parte alia imperii gestas diu reliquis ignotas fuisse.

Diebus nostris, electromagneticis undis rerum cognitiones et

Franco Maceri, Preside della Facoltà di Ingegneria presso l'Università di Tor Vergata, porge il benvenuto ai presenti.

notitiae quocumque vehuntur.

Undarum congeries tamen saepe inter se permiscetur, ita ut cognitionum rationis pars pereat, sicut efflatus vocis in strepitu et clamore.

Per gentile concessione del Ministero per i Beni Culturali e Ambientali

La via Appia in prossimità del mausoleo di Cecilia Metella. Sullo sfondo l'agro romano dove oggi sorge l'Università di Tor Vergata (da una stampa di François Morel; Istituto Nazionale per la Grafica, Roma).

Eminus ergo vi electrica saepe nuntiatur per numerorum continuationem seriemque ita ut, quamvis longo interiecto intervallo, quovis deferantur et percipiantur notitiae, minime errore perturbatae.

Qua via undae in numeros optime convertantur, plures, copia animi fervidique ingenii ornati, doctissime investigaverunt.

Ex illis caput extollit Andreas Viterbi qui, suo excogitato algorismo, novas computandi atque eminus nuntiandi ope machinae in inani sitae rationes protulit, qua de causa illum vero artis suae conditorem, una cum Wilhelmo Marconi, Henrico Nyquist et Claudio Shannon habere possumus.

Nunc igitur Aldus Roveri, Professor Clarissimus, vir summa ornatus doctrina, qui multos per annos apud utrumque Urbis Studium innumeros discipulos sapientia locupletavit, candidati nostri facta laudibus ornabit et suo celebrabit sermone.

I Professori intervenuti festeggiano Viterbi al termine della cerimonia.

Conferimento laurea Honoris Causa

Elogio del Candidato

Aldo Roveri: Autorità, illustri Ospiti, cari Colleghi e Studenti, Signore e Signori, ho avuto il compito di pronunciare l'Elogio del Candidato nel conferimento della laurea Honoris Causa in Ingegneria delle Telecomunicazioni a Andrew J. Viterbi. Di questo compito sono grato ai colleghi e amici dell'Università degli Studi di Roma "Tor Vergata". Sono infatti onorato di parlarvi di una personalità di eccezionale rilevanza nel quadro dello sviluppo della teoria e della tecnica delle comunicazioni negli ultimi trent'anni. Eccezionali sono infatti i suoi contributi all'avanzamento delle conoscenze nel campo delle comunicazioni numeriche, così come di straordinario rilievo è la articolazione della sua vita professionale come Ricercatore universitario, come Progettista nello sviluppo industriale di sistemi tecnicamente di assoluta avanguardia e come Imprenditore nel dar vita ad iniziative industriali di pieno successo.

Il Professor Viterbi è nato a Bergamo nel 1935. Il suo curriculum di studi si è svolto dapprima presso il Massachusetts Institute of Technology (MIT) e successivamente presso la University of Southern California, ove ha conseguito il dottorato di ricerca nel 1962.

Nel suo primo impiego dopo aver conseguito i B.S. e M.S. degrees presso il MIT (1957), ha fatto parte del gruppo di progetto, presso il C.I.T. del Jet Propulsion Laboratory, che ha progettato e realizzato il sistema di telemetria sull'Explorer I, il primo satellite USA con successo. Nei primi anni Sessanta presso lo stesso laboratorio, egli è stato uno dei primi a riconoscere le potenzialità e a proporre l'impiego delle tecniche di trasmissione numerica per sistemi di telecomunicazioni spaziali e via satellite. Può quindi definirsi un vero e proprio pioniere delle tecnologie di telecomunicazione che si sono affermate circa dieci anni dopo e che oggi stanno guidando i più recenti sviluppi della comunicazione a distanza.

Come docente presso la School of Engineering and Applied Science della UCLA (University of California Los Angeles) dal 1963 al 1973, il Prof. Viterbi ha svolto studi di fondamentale importanza nella teoria delle comunicazioni numeriche, pubblicando numerosi articoli e due libri. Per questa attività ha ricevuto numerosi premi e riconoscimenti in ambito sia nazionale che internazionale: di questi darò un cenno nel seguito.

L'intervento di Aldo Roveri, Professore presso l'Università di Roma "La Sapienza".

Motivazione del premio "Award Alexander Graham Bell" consegnato nel 1984 a Viterbi dall'IEEE (Institute of Electrical and Electronics Engineers).

Andrew J. Viterbi

1984 Alexander Graham Bell Medal
"For fundamental contributions to telecommunication theory and practice and for leadership in teaching."

Andrew J. Viterbi was born on March 9, 1935 in Bergamo, Italy. Arriving in the U.S. in 1939, he received the B.S. and M.S. degrees in Electrical Engineering in 1957 from MIT and the Ph.D. in Electrical Engineering in 1962 from the University of Southern California.

In his first employment after graduating from MIT he was a member of the project team at CIT Jet Propulsion Laboratory which designed and implemented the telemetry equipment on the first successful U.S. satellite, Explorer I. In the early sixties at the same laboratory, he was one of the first communication engineers to recognize the potential, analyze and propose digital transmission techniques for space and satellite telecommunication systems.

As a professor in the UCLA School of Engineering and Applied Science from 1963 to 1973, Dr. Viterbi did fundamental work in digital communication theory, and authored numerous research papers culminating in two books on the subject. The first, on phase coherent communication techniques, was the first comprehensive research monograph on the phase locked loop and its application to both tracking and demodulation. The second (jointly authored), on coding and information theory, contained the work for which he is best known: the Viterbi algorithm for maximum likelihood decoding of convolutional codes, which also found application in a host of other digital demodulation and processing applications. This concept has deeply affected information and communication theory, both for its tutorial elegance and for its applications to a broad class of communication problems.

The practical development of these theoretical principles led to the founding of LINKABIT Corporation, together with Dr. Irwin Jacobs. Dr. Viterbi was Executive Vice President of LINKABIT from 1974 to 1982. Since 1982 he has been President of M/A-COM LINKABIT, Inc.

Dr. Viterbi is a member of the U.S. National Academy of Engineering and a Fellow of IEEE. He is past Chairman of the Visiting Committee for the Electrical Engineering Department of Technion, Israel Institute of Technology, and he is presently a member of the MIT Corporation Visiting Committee for Electrical Engineering and Computer Science. He is also Chairman of U.S. Commission C of the International Radio Scientific Union (URSI) and a past member of the Army Science Board. He has been active at various times as a member or chairman of the Board of Governors of the IEEE Information Theory Group as its Transactions Associate Editor for coding. Since 1975, he has been Adjunct Professor of Electrical Engineering and Computer Science at the University of California, San Diego.

Previous recognition includes three paper awards, culminating in the 1968 IEEE Information Theory Group Outstanding Paper Award. He has received two other major society awards: the 1975 Christopher Columbus International Award (from the Italian National Research Council, endowed by the City of Genoa); and jointly with Irwin Jacobs, the 1980 Aerospace Communications Award (from AIAA).

In spite of his current corporate administrative duties, he has managed to remain technically current, having recently proposed new spread spectrum processing techniques for jam resistant communication and for digital mobile radio.

Andrew and his wife, Erna, have three children: Audrey, a doctoral candidate in the EECS Department at UC Berkeley; Alan, a Coro Foundation Fellow, pursuing a public policy graduate program; and Alexander, a seventh grader. As a family, they enjoy travelling together.

Tra i numerosi contributi di particolare rilevanza scientifica in questo periodo accademico, mi limito ad accennare al suo lavoro nel campo della codifica convoluzionale. I risultati sono riportati in un celebre articolo pubblicato nell'aprile 1967 sulle IEEE Transactions on Information Theory.

In questo articolo viene proposta una nuova strategia di decodifica convoluzionale, che è basata su un metodo di ricerca a traliccio e che, da allora, è universalmente nota come "Algoritmo di Viterbi". Questo, grazie all'impiego di tecnologie VLSI e delle potenzialità di elaborazione offerte dai microprocessori, può essere realizzato a basso costo pur con prestazioni assai vicine a quelle teoriche.

Nel 1968 è stato co-fondatore della LINKABIT Corporation, dove, dal 1974 al 1982, ha operato come Vice Presidente Esecutivo e, dal 1982 al 1984, come Presidente. La LINKABIT è una società di apparati di telecomunicazioni per applicazioni spaziali: con il suo contributo decisivo, ha acquistato una posizione di assoluta preminenza in ambito mondiale nella progettazione e nella realizzazione di apparati ricetrasmittenti (modemodulazione e codecodifica) di avanguardia. In questo periodo ha fornito importanti contributi nella definizione di architetture modem in tecnica numerica e nel settore dei metodi di sincronizzazione.

Andrew J. Viterbi (terzo da sinistra) insignito del Dottorato Onorario in Ingegneria dalla "University of Waterloo" (Canada, 1990).

Nel 1985 unitamente ad altri, ha fondato la QUALCOMM Incorporated, una società specializzata nei sistemi di comunicazione mobile via terra e via satellite e nelle tecniche di elaborazione del segnale. In questa ulteriore attività imprenditoriale, ove occupa le responsabilità di Vice Chairman e di Chief Technical Officer, è di grande rilievo la sua opera nello sviluppo di un sistema radiomobile con tecnica di accesso a divisione di codice (CDMA), che è attualmente alla base dello standard USA per i prodotti cellulari in tecnica numerica.

A testimonianza della universale considerazione che il mondo tecnico scientifico ha manifestato e tuttora manifesta nei confronti del Prof. Viterbi, citerò alcuni dei riconoscimenti e dei premi da lui ricevuti:

- *è membro della U.S. National Academy of Engineering ;*
- *ha presieduto la Visiting Committee per l'Electrical Engineering Department del Technion-Israel of Technology;*
- *è stato Distinguished Lecturer all'University of Illinois e all'University of British Columbia;*
- *è stato membro della Visiting Committee per l'Electrical Engineering and Computer Science presso il MIT;*

Un altro riconoscimento internazionale conferito a Viterbi: il premio della Fondazione NEC C&C (Tokyo, 1992).

アンドリュウ J.ビタービ 博士

Chief Technical Officer
Qualcomm Incorporated

外惑星探査衛星ボイジャーⅠ・Ⅱ号、及び金星探査衛星マゼラン号の輝かしい成功を可能にした深宇宙ディジタル通信システムの設計と開発に対する基本的貢献

歴と主なる業績

アンドリュウJ・ビタービ博士は、1935年3日、イタリー・ベルガモに生まる。1939年、に移民。1957年、ＭＩＴより、電気工学のおよびS.M.の学位を修得、次いで1962年、リフォルニア大学より、電気工学のPh.D.を修得された。

年、同博士は、カリフォルニア工科大学ト推進研究所の通信研究部に入所された。部の長は、当初はエバーハルト・レクティン博士で、次いでウォルター・ビクター氏に代われた。共に今回の連名受賞者である。そこでは、米国最初の衛星および初期の惑星探査ミッションの為のテレメトリー・システムの開発チームのメンバーとして、Phase-Locked Loops,デジタル変調法および符号化問題に関する理論の研究に寄与された。この初期における経験は、同博士の全経歴の方向付けに大きく影響を与えたのである。

●1963年、ビタービ博士は、ＵＣＬＡの工学部および応用物理部の教授に就任され、次の十年間は、主としてディジタル通信に関する教育と研究に従事された。その間、２冊の著書と多数の論文を著わされた。その幾つかは受賞された。特記すべきことは、この期間に、同博士は、Convolutional Codesの復号法に関するアルゴリズムを提案されたことである。そしてこれは、本プロジェクトおよび電気通信や記録技術分野に広く応用された。

●同博士は、1968年、LINKABIT Corporationの共同設立者として入社され、1973年にはExecutive Vice Presidentに、次いで、1982年にはPresident and CEOに就任された。大学教授連によって発足したこの起業家活動は、1985年には、社員数1500名を超える会社にまで成長した。そしてそれは、事前誤り訂正符号法、マイクロプロセッサベースの衛星用モデム、Ku－band VSAT's及びデータ伝送と衛星画像放送の保護の為の商用暗号方式等の製品とシステムの開発のパイオニアとなった。

●1985年、上記会社を退職３ヶ月後、ビタービ博士は、イルウィン ヤコブス博士と共同で、第二の会社、QUALCOMM, Inc. を設立させた。そこでは彼は、副会長兼主任技師長として、衛星と地上伝送技術を用いて、移動体および個人通信用の新しいシステムとサービスの開発に専念された。その最初の製品とサービスは、交通産業のための広域の衛星ベースの通信と位置確認網の提供と活動であった。それは現在、北米および欧州において、３万ヶ所以上のモバイルプラットフォームでサービスを提供している。この会社の主要技術力となったのは、分散スペクトル符号分割多重アクセス法を採用した所の小地区およびパーソナルに対する同時普遍的な音声データ通信サービスの創造にある。この技術は、まだ世界的普及とまでは行っていないがしかし、ビタービ博士にとっては、このことは彼の経歴のスタートであるＪＰＬに於いて、彼が学び且つ仕上げた理論の自然的発展の成果ということが出来ると思われる。

主たる受賞

- IEEE Information Theory Outanding Paper Award (1967)
- Christopher Columbus International Award and Medal (1975)
- IEEE Alexander Graham Bell Award and Medal (1984)
- Marconi International Fellowship Award (1990)
- IEEE Information Theory Society Shannon Lecturer(1991)

ビタービ博士は、諸外国の有名大学において特別招待講演を行なっておられる。その代表的な所を挙げれば、次の通り：南カリフォルニア大学で有名工学卒業生として（1986)、カナダ、ウォーターロー大学で、名誉工学博士号受領に当たって(1990)、その他Harvey Mudd College (カリフォルニア・クレアモント)、MIT,. Technion (Haifa, Israel)等。更に、幾つかの米国政府の諮問委員を務められている。また、IEEEのFellow (1973)、および米国のNational Academy of Engineeringの会員(1978)にも選ばれておられる。

ビタービ博士は、1963年以来、カリフォルニア大学組織と関係を続けておられ、現在、同大学サンディエゴ校の電気・コンピュータ工学部の教授をしておられる。

同博士は、その教育者、研究者、そして企業家としての諸活動に満足されておられることに加えて、エルナ夫人、三人の子供さん、心待ちにされているお孫さんの誕生など、幸福な家庭にも恵まれておられる。

- *ha ricevuto (1986) il Premio "Annual Outstanding Engineering Graduate" da parte della University of Southern California;*
- *è stato insignito di un Dottorato Onorario in Ingegneria dall'University of Waterloo;*
- *ha presentato (1991) la Shannon Lecture all'International Symposium on Information Theory;*
- *ha ricevuto premi per articoli pubblicati quali l'IEEE Information Theory Group Outstanding Paper Award nel 1968 e lo Stephen O. Rice Award (in collaborazione) nel 1994;*
- *ha ulteriormente ricevuto nel 1975 il premio internazionale Cristoforo Colombo assegnato dal Consiglio Nazionale delle Ricerche per iniziativa della Città di Genova;*
- *gli sono stati anche assegnati nel 1994 la IEEE Alexander Graham Bell Medal e nel 1990 il Marconi International Fellowship Award;*
- *è stato infine coassegnatario nel 1992 del NEC C&C Foundation Award e nel 1994 del Eduard Rhein Foundation Award per la ricerca di base.*

In conclusione, posso affermare che Andrew J. Viterbi, per la sua produzione scientifica e per gli eccezionali contributi offerti in campo industriale, occupa una posizione di assoluta preminenza nella storia delle telecomunicazioni, a fianco di nomi di grande prestigio quali Guglielmo Marconi, Harry Nyquist e Claude E. Shannon.

Viterbi firma il registro del premio "16th Marconi International Fellowship Award". A destra la Signora Gioia Marconi Braga (1990).

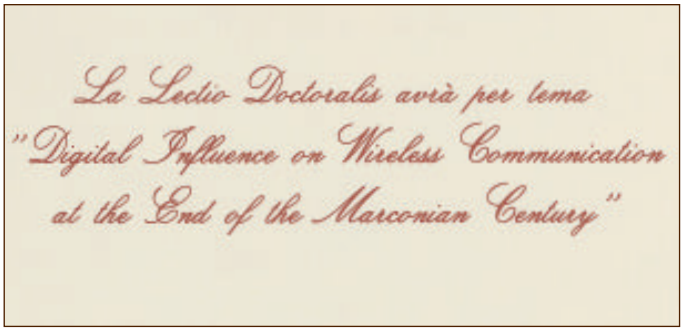
Condivido quindi pienamente la decisione di questo Ateneo di esprimere il proprio riconoscimento a una persona con meriti professionali di così spiccata rilevanza. Al riguardo rilevo che questo conferimento di Laurea Honoris Causa ha due peculiarità. Una prima, di valenza generale, è legata alla vocazione di questa sede accademica, manifestata fin dall'inizio della sua ancora breve vita, a collocarsi in un contesto internazionale, stabilendo tra l'altro, come in questo caso, ponti culturali con persone di alto livello che siano di riferimento scientifico per il proprio ambiente di ricerca.

Una seconda peculiarità, di carattere più specifico riguardante la specializzazione della Laurea, si ricollega alla tradizione della cultura tecnica in questo Paese, ove le telecomunicazioni sono state seguite, fin dalla loro nascita, con impegno di studi e di iniziative industriali e ove, più in particolare, la transizione dalle tecnologie analogiche a quelle numeriche, e cioè la fase evolutiva che è stata illuminata dall'opera del Prof. Viterbi, è stata governata, presso varie sedi universitarie, con la formazione di scuole che hanno acquisito elevato prestigio internazionale e, presso società manifatturiere e di esercizio, con la realizzazione di sistemi, di infrastrutture e di servizi che, pur tra i tanti problemi che affliggono il sistema-Paese, ci collocano tra le realtà tecnologicamente più avanzate verso l'attuazione dell'era dell'informazione.

Conferimento laurea Honoris Causa

Invito a presentare la Lectio Doctoralis

Facultatis Decanus: Domine Andreas Viterbi, dicas nobis, quaeso, dissertationem tuam.



Andrew J. Viterbi: Professor Finazzi Agrò, Magnifico Rettore; Professor Maceri, Preside della Facoltà di Ingegneria; Professor Valdoni, Presidente del Consiglio di Corso di Laurea in Ingegneria delle Telecomunicazioni; Professor Roveri, dell'Università di Roma "La Sapienza"; Ingegner De Guzzis, Presidente dell'Associazione Italiana degli Ingegneri delle Telecomunicazioni; Ingegner de Julio, Ingegner Vannucchi, Illustri Ospiti, Colleghi, Signore e Signori,

sono veramente commosso per l'elogio e per il grande onore che mi è stato conferito.

Con il vostro permesso, vorrei dedicare la mia "Lectio Doctoralis" alla memoria di mio padre, Professor Achille Viterbi.

Fu il mio primo e costante maestro.

Tra le tante cose che mi insegnò vi erano il latino - purtroppo ormai quasi dimenticato - e l'italiano, che sarà sempre parte di me.

Mio padre amò questa sua Patria tutta la vita, malgrado gli eventi sfortunati che lo costrinsero a lasciarla.

La gratificazione che è stata conferita a suo figlio onora soprattutto la sua memoria.

Lectio Doctoralis

Digital Influence on Wireless Communication at the End of the Marconian Century

In keeping with the theme of this morning's seminar, I have chosen to speak on the topic of technology and industrial policy as it impacts the current rapid evolution of digital wireless.

The wireless industry is celebrating its centennial. It was Guglielmo Marconi at the dawn of the twentieth century who, building on the scientific theories and experiments of nineteenth century physicists, discovered the practical means of achieving wireless electromagnetic transmission and reception. With a vision and courage well beyond his contemporaries, he almost single-handedly created an industry whose ultimate impact and benefits were beyond the comprehension of his age. In the subsequent decades, with the benefit of evolving technology, wireless emerged from a means for seeking emergency assistance by sinking ships to its wider role for universal broadcast of news and entertainment. The Second World War created the crisis environment which brought the best scientific and technical talents to the service of rapidly exploiting wireless technology. Out of this came improved radar, secure and jam-proof communication, electronic control and the primitive beginnings of digital computation. Equally important for the future evolution of the field in the United States, an industrial policy emerged, driven by the perceived need for defense preparedness, which was solidly based on scientific research and development. The name generally attributed to this policy is that of Vannevar Bush, a Professor at MIT and a pioneer in electronic computation. The Post-war committee he chaired, on behalf of the Department of Defense, empowered several defense agencies to invest tax dollars in both funda-

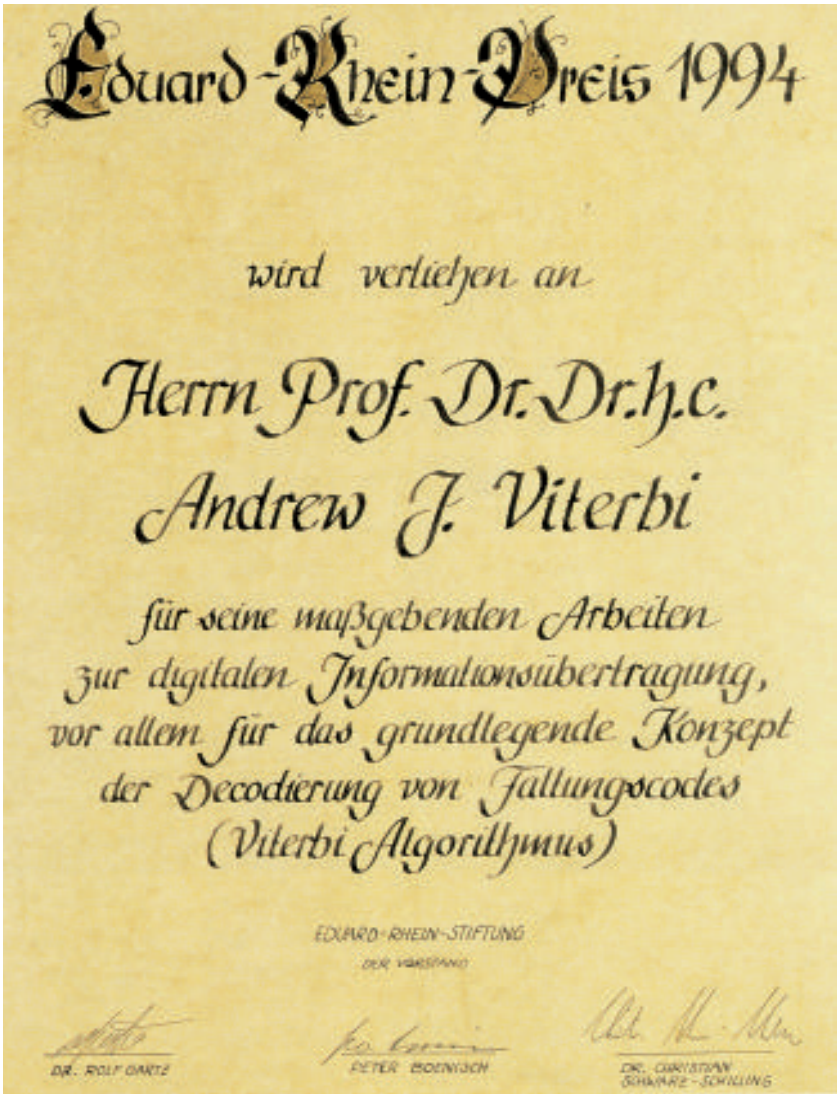
In sintonia con il tema del seminario di questa mattina, ho scelto di parlare della tecnologia e della politica industriale poiché essa influenza l'attuale rapida evoluzione delle radiocomunicazioni numeriche.

L'industria delle radiocomunicazioni celebra quest'anno il centenario. Fu Guglielmo Marconi che all'inizio del Ventesimo secolo, sulla scorta delle teorie scientifiche e degli esperimenti dei fisici del Diciannovesimo secolo, scoprì i mezzi con i quali realizzare praticamente la trasmissione e la ricezione senza fili delle onde elettromagnetiche. Con un intuito e un coraggio di gran lunga superiori a quelli dei suoi contemporanei, creò quasi da solo un'industria il cui impatto estremo ed i cui benefici non erano comprensibili ai suoi tempi.

Nei decenni successivi, con i vantaggi offerti dall'evoluzione della tecnologia, le radiocomunicazioni non furono impiegate solo come mezzo di emergenza usato dalle navi in pericolo per chiedere soccorso ma assursero anche al ruolo più ampio di mezzo per la diffusione universale delle notizie e per l'intrattenimento. La seconda guerra mondiale creò l'atmosfera di crisi che portò i migliori talenti scientifici e tecnici ad assumersi il compito di sfruttare rapidamente la tecnologia delle radiocomunicazioni. Si ebbero così il miglioramento del radar, le comunicazioni protette ed esenti da interferenze intenzionali, il controllo elettronico e i primordi del calcolo numerico. Di eguale importanza per l'evoluzione successiva del settore negli USA, emerse una politica industriale pilotata dalla sentita esigenza di una preparazione alla difesa, fondata su solide basi di ricerca e sviluppo. Il nome generalmente associato a questa politica è quello di Vannevar Bush, professore del MIT e

Il Rettore Alessandro Finazzi Agrò invita Andrew J. Viterbi a tenere la Lectio Doctoralis.

mental and applied research, much of which was communication related. The research contributions of the Office of Naval Research, Air Force Office of Scientific Research, Army Research Office and most importantly the later constituted ARPA (Advanced Research Projects Agency) are widely recognized. Much of this research was performed by academic institutions or their specially created research arms, but an increasing fraction of the funding also went to industrial concerns, with small entrepreneurial companies often receiving preference. These R&D funds, which constituted only a minimal fraction of the total defense appropriation for weapons systems, often had a disproportionate impact on industrial



Premio della Fondazione "Eduard Rhein" per la ricerca di base (Amburgo, 1994).

developments. Which is not to say that all United States R&D was government funded. In fact, the seed research for digital wireless came out of the Bell Telephone Laboratories, totally funded by the monopolistic Bell Telephone System, later called AT&T. In the very early postwar years, from two unrelated research groups at Bell Labs in Murray Hill, New Jersey - led respectively by William Shockley and

pioniere del calcolo elettronico. Il comitato costituito al termine degli eventi bellici e da lui presieduto, per conto del Ministero della Difesa, autorizzò varie agenzie per la difesa ad investire i fondi dei contribuenti nella ricerca sia di base che applicata, in gran parte orientata alle comunicazioni. I contributi della ricerca di Enti quali l'*Office of Naval Research*, l'*Air Force Office of Scientific Research*, l'*Army Research Office* e soprattutto il più recente *ARPA (Advanced Research Projects Agency)* sono ampiamente riconosciuti. Questa attività di ricerca fu condotta in larga misura da istituzioni accademiche o da enti di ricerca da esse appositamente costituiti, ma una parte crescente dei finanziamenti fu pure assegnata ad aziende manifatturiere, privilegiando in molti casi piccole realtà imprenditoriali. Questi fondi di ricerca e sviluppo, che costituivano solo una piccola frazione del totale dei fondi accantonati dalla difesa per i sistemi d'arma, ebbero in molti casi un impatto molto elevato sullo sviluppo industriale. Ciò non significa affatto che tutta l'attività di ricerca e sviluppo negli Stati Uniti era finanziata dal Governo. Infatti, la ricerca germinale delle radiocomunicazioni numeriche scaturì dai Bell Telephone Laboratories, finanziati completamente dal gruppo monopolistico Bell Telephone System, divenuto successivamente AT&T. Nei primissimi anni del dopoguerra, da due gruppi di ricerca non collegati dei Laboratori Bell di Murray Hill nel New Jersey - diretti rispettivamente da William Shockley e da Claude Shannon - scaturirono i principi fondamentali dell'elettronica dello stato solido, che si concretizzarono nel transistor, e il concetto di efficienza massima nell'elaborazione e trasmissione delle informazioni numeriche. Conviene osservare che nonostante i Laboratori Bell fossero un'organizzazione commerciale, il propulsore di questa ricerca fu l'esigenza bellica; nel caso di Shannon, la teoria dell'informazione provenne dal suo celeberrimo studio sulla "Teoria matematica della segretezza", ora non compreso più nella lista dei documenti riservati.

Nonostante il notevole sforzo di ricerca, l'impatto sulle telecomunicazioni commerciali e di consumo fu modesto per buona parte dei primi tre decenni del dopoguerra. Vero è che le trasmissioni televisive analogiche gradualmente giunsero a maturazione, divennero un servizio quasi universale e permisero l'introduzione del colore; la commutazione elettronica sostituì le centrali telefoniche elettromeccaniche e gli elaboratori numerici divennero più veloci con molta più memoria e dotati di programmi con accresciute potenzialità in

Claude Shannon - came the fundamental concepts of solid state electronics, embodied in the transistor, and of maximally efficient digital information processing and transmission. It is noteworthy that though Bell Labs was a commercial entity, the impetus for this research came from wartime research; in Shannon's case, information theory came from his now unclassified and celebrated "Mathematical Theory of Secrecy."

In spite of all the momentum, the impact on commercial and consumer telecommunications was modest over most of the first three postwar decades. True, analog television broadcasting matured, became an almost universal service and converted to color; electronic switching replaced mechanical telephone exchanges, and digital computers became faster with far greater memory and with improved software capabilities to facilitate their use. With the exception of television, however, these advances were invisible to the general public. Mainframe computers, for example, initially were of benefit solely to academic researchers and to the largest businesses.

Andrew J. Viterbi assieme, da sinistra, a Irving Reed e Gustave Solomon (inventori del Codice Reed-Solomon) ed a Ray Bradbury (scrittore di fantascienza) presso il Jet Propulsion Laboratory nel corso dell'incontro del Voyager con il pianeta Urano (1983).

In short, the roots of digital wireless research took nearly a generation to mature to the level of practical applications. An important catalyst was the communication satellite. With the incentive created by the launching of Sputnik, the U.S. defense establishment again mounted an immense program to overcome Soviet supremacy in Space. Within less than a decade the first commercial communication satellites became operational and in the seventies they began to convert to digital transmission. The first commercial digital wireless initiative was driven by the cost efficiencies achieved through digital transmission, as well as compression, produced by lowering the required receiver signal-to-noise ratio, and hence the required transmitter power which results in reduced satellite weight in orbit. Conversely, for a given fixed weight, either transmission rate can be increased or receiver antenna size reduced. Today these benefits of digital wireless are available to most business users of VSAT's (Very Small Aperture Terminals), both for fixed and mobile service, and increasingly to consumers with digital Direct Broadcast Satellite receivers.

Yet, the evolution of digital wireless has been much more dramatically impacted by the establishment in the late sixties and seventies of a number of small entrepreneurial corporations, mostly on the west coast of the United States. The earliest of these were started by solid state physicists and engineers who came from premier research and academic institutions and recognized that not only speed, size and weight considerations drive the need

modo da facilitarne l'impiego. Tuttavia, eccezion fatta per la televisione, questi progressi non furono percepiti dal grande pubblico. Gli elaboratori centrali, ad esempio, erano usati all'inizio esclusivamente dai ricercatori delle Università e dalle grandi imprese.

In breve, per passare al livello delle applicazioni pratiche, la ricerca nelle radiocomunicazioni numeriche impiegò quasi una generazione. Il satellite per comunicazioni fu un importante catalizzatore. Con l'incentivo creato dal lancio dello Sputnik, il settore della difesa degli Stati Uniti d'America avviò un altro programma di enorme impegno per superare la supremazia sovietica nello spazio. Nell'arco di meno di un decennio furono messi in servizio i primi satelliti per comunicazioni commerciali che negli anni Settanta cominciarono a convertirsi alle tecniche di trasmissione numerica. La prima iniziativa in materia di radiotrasmissione numerica commerciale ha ricevuto un impulso dai risparmi di costo realizzati con la trasmissione numerica, nonché mediante la compressione, ottenuta abbassando il rapporto segnale-rumore richiesto in ricezione e quindi la potenza di trasmissione necessaria, che a sua volta riduce il peso del satellite in orbita. Fissando al contrario un certo peso, è possibile aumentare il ritmo di trasmissione oppure ridurre la dimensione dell'antenna ricevente. Oggi questi vantaggi delle radiotrasmissioni numeriche sono disponibili a gran parte degli utenti di sistemi VSAT (terminali dotati di antenna con apertura piccola) per servizi sia fissi che mobili, e anche a un numero crescente di clienti dotati di ricevitori per radiodiffusione numerica diretta da satellite.

L'evoluzione delle radiotrasmissioni numeriche è stata però influenzata molto di più dalla nascita negli ultimi anni Sessanta e negli anni Settanta di alcune organizzazioni imprenditoriali di piccole dimensioni, localizzate principalmente sulla costa occidentale degli Stati Uniti. Le prime tra queste sono state fondate da fisici e tecnici dello stato solido provenienti da isti-

for large scale integration (LSI) but that cost reductions of orders of magnitude follow and dominate all other considerations. First with memories and later with microprocessors, these pioneers changed forever the landscape of the computer and communication industries, and indeed all of electronics. In short, they liberated circuit and system designers from the constraints of specific physical devices, replacing the classical analog electronic circuit by a powerful mathematical processor whose potential capabilities (high speed, large memory, small size and high degree of integration) grow exponentially by doubling approximately every eighteen months. This feature, known as Moore's Law, after Gordon Moore of Intel, replaces a pragmatic and simplistic circuit implementation by the realization of the best algorithm to perform the task, whether that be for a wireless modem, digital television receiver, an optimal control system or a sophisticated flight simulator.

But before this current state of our technology became possible there was an intervening chapter of the story, namely the rendering of the all-powerful and intimidating digital computer mainframe into a consumer product, residing in virtually every office in the industrialized world as well as an increasing percentage of homes. This feat was accomplished

Viterbi tiene la Lectio Doctoralis prima del conferimento della laurea Honoris Causa in Ingegneria delle Telecomunicazioni.

not by the experienced scientists and engineers of the integrated solid state electronics revolution, but by a totally different class of technological pioneers, mostly bright, through uneducated, yet highly enterprising university drop-outs, who in the seventies were among the first to recognize the low-cost potential of solid state integration. Riding on the coat-tails of their erudite elders, these upstarts used their tools and building blocks to produce the hardware and the software embodiments of the first personal computers. While some skeptics regarded these as little more than sophisticated toys, the giant IBM Corporation caught a glimmer of the future possibilities and both supported and competed with

tuti accademici e di ricerca di primo piano. Essi riconobbero che non solo considerazioni relative a velocità, dimensioni e peso portano all'esigenza di integrazione su larga scala (LSI), ma anche che con questo impiego si ottiene una riduzione dei costi di ordini di grandezze, che finisce per dominare tutte le altre considerazioni. Prima con le memorie e successivamente con i microprocessori, questi pionieri hanno cambiato radicalmente il panorama dell'industria dell'informatica e delle telecomunicazioni, e in effetti di tutto il settore dell'elettronica. In breve, essi hanno liberato i progettisti di circuiti e di sistemi dai vincoli legati all'uso di dispositivi fisici specifici, sostituendo il classico circuito elettronico analogico con un potente processore matematico le cui capacità potenziali (alta velocità, grande memoria, piccole dimensioni ed elevato grado di integrazione) crescono in modo esponenziale, con un raddoppio ogni 18 mesi circa. Questa caratteristica, evidenziata da Gordon Moore della Intel, e perciò nota come legge di Moore, sostituisce la realizzazione di circuito pragmatica e semplicistica con quella del miglior algoritmo in grado di svolgere il compito assegnato, sia che si tratti di un modem per radiotrasmissioni, di un ricevitore di televisione numerica, di un sistema di controllo ottimo o di un sofisticato simulatore di volo.

Prima però di pervenire alla fase presente della tecnologia si ebbe un capitolo intermedio della storia, e cioè la conversione del grosso elaboratore centrale, potentissimo e temuto, in prodotto di consumo presente praticamente in ogni ufficio del mondo industrializzato e in misura crescente anche nell'ambito domestico. Questa impresa fu compiuta non da scienziati e tecnici esperti della rivoluzione dell'elettronica integrata a stato solido, ma da una classe totalmente diversa di pionieri della tecnologia, in gran parte universitari falliti molto intelligenti e altamente intraprendenti, anche se di livello di istruzione modesto, che negli anni Settanta furono tra i primi a intravedere la potenziale riduzione di costo permessa dall'integrazione dei semiconduttori. Operando sulla scia dei loro predecessori eruditi, questi personaggi della "gavetta" hanno usato i loro "attrezzi" e i loro "mattoni" per produrre i primi campioni di hardware e di software dei personal computer. Mentre alcuni scettici li consideravano poco più di giocattoli sofisticati, il gigante IBM si rese subito conto delle future possibilità, e partecipò attivamente sia offrendo appoggio che facendo concorrenza a questo manipolo di giovani lungimiranti per dare all'industria un'impronta di rispettabilità e successivamente di supremazia. Oggi i computer da tavolo e le work stations hanno praticamente assorbito il precedente segmento inferiore dell'industria dei grandi calcolatori,

this visionary band of youngsters to bring the industry to respectability and ultimately supremacy. Today's desktop computers and workstations have virtually dismantled the former low-end of the mainframe industry, the minicomputer which had been the market leader in the seventies and early eighties. Only the supercomputer seems likely to survive the challenge of the ever more powerful desktop work stations, but with a very limited market, and hence dependence on the support of government or of very well financed specialized industries such as investment institutions and cinema producers.

The impact of the personal computer revolution on wireless telecommunication has been to accelerate the pace of integration, speed and cost reduction required to implement the sophisticated optimal algorithms for digital cellular telephony, digital broadcast home satellite and terrestrial receivers and wireless modems for personal computers, all at a cost of a few hundred dollars or less. The power of the personal computer has created one other major telecommunication phenomenon, the INTERNET, predominantly wireline but already beginning to be accessible by wireless means. Begun in the early seventies as the U.S. Defense Department ARPANET, it was designed to interconnect universities and research institutes for the purpose of sharing software resources and messages for collaborative research purposes. We all know how this has grown, almost like a living organism, feeding on the fruits of the latest technology, to reach the general public in virtually all corners of our planet. Enhanced by the World Wide Web software structures developed at Europe's CERN, it has become an international resource library, as yet somewhat chaotic, but becoming increasingly organized and useful as "Search Engines" improve and become more powerful and intelligent.

So what can we derive from this fascinating but complex story to guide us in the future? I propose to split this question into two parts, asking first what are the present challenges and opportunities based on our current state of digital telecommunication technology? And secondly, what are the economic, political and societal structures which enhance or which constrain our collective abilities to achieve these goals? The first question has a relatively direct answer, the second invites a multifaceted response.

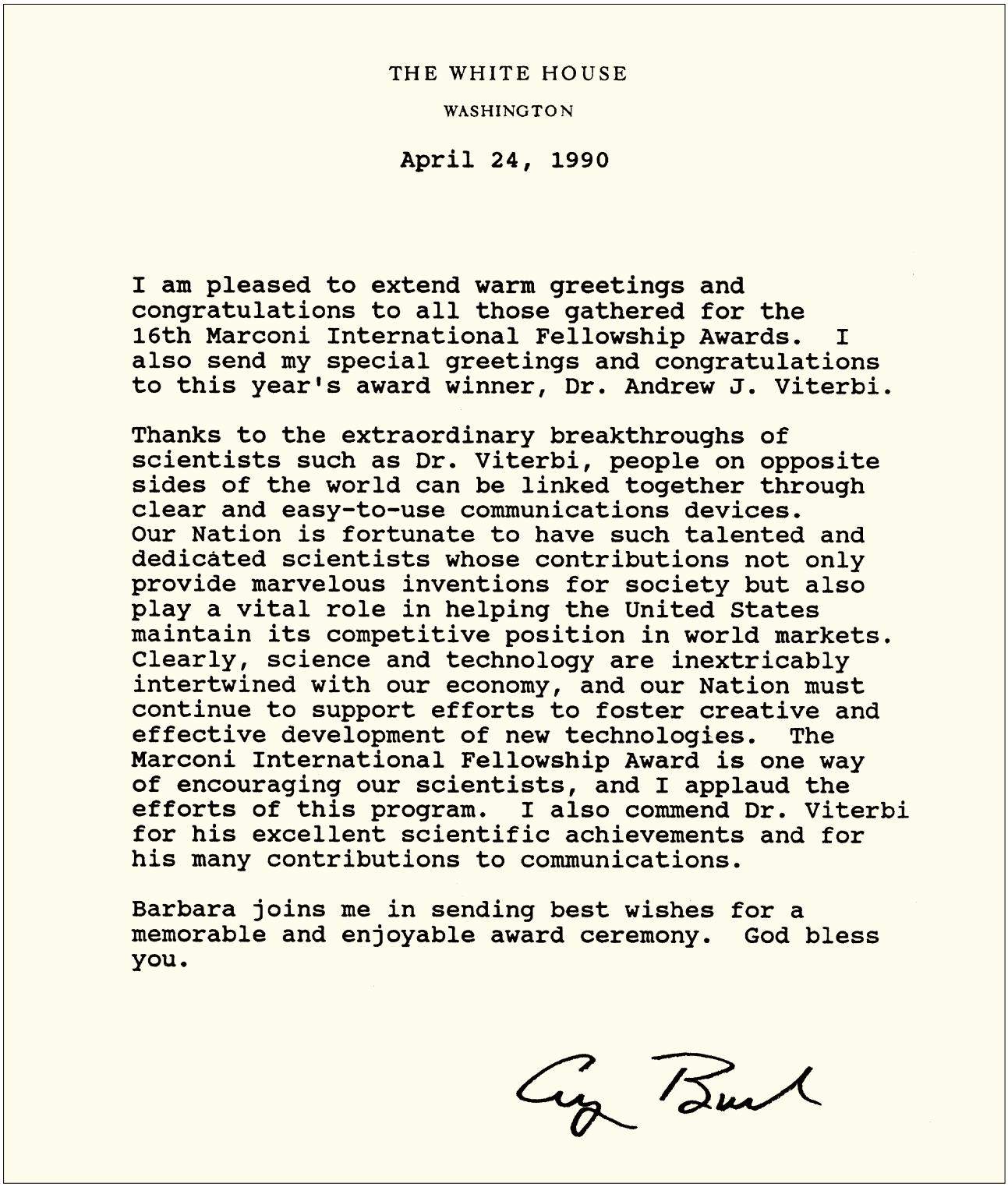
My first answer can be summarized as two perceived needs, for which the market appears assured: "Interactive Data, Knowledge and Pastimes at the click of a (PC) mouse;" and "Anywhere, Anytime Personal Communication." As simplistic and obvious as these two statements may appear, they both present complex societal implications. For the first, currently in the U.S. there is a bitter unresolved conflict between the television manufacturers and broadcasters, on one hand, and the PC industry on the other, over digital high definition television standards (more on this later). Beyond the competitive advantage that each seeks, there are serious implications. While there is every likelihood that entertainment and information will somehow merge, both becoming interactive, does the general public really want to participate actively

in particolare il minicomputer che era leader di mercato tra gli anni Settanta e i primi anni Ottanta. Solo il supercomputer sembra oggi in grado di sopravvivere alla sfida delle sempre più potenti stazioni di lavoro da scrivania, ma con quote di mercato assai limitate, e quindi con una notevole dipendenza dagli aiuti governativi o da industrie specializzate con larghe disponibilità finanziarie come gli investitori istituzionali e i produttori cinematografici.

La rivoluzione del personal computer ha influenzato le radiocomunicazioni facendo accelerare il ritmo di integrazione, la velocità e la riduzione dei costi necessari per introdurre sofisticati algoritmi ottimali per la telefonia cellulare numerica, i satelliti per trasmissioni numeriche terminate direttamente presso l'utenza (e i relativi ricevitori terrestri) nonché modem senza fili per personal computer, tutti a un costo non superiore a poche centinaia di dollari. La potenza del personal computer ha fatto nascere un altro grande fenomeno nelle telecomunicazioni, INTERNET, prevalentemente basato su collegamenti cablati ma già accessibili anche via radio. Avviata agli inizi degli anni Settanta nell'ambito del programma ARPANET del Ministero della Difesa USA, la rete era stata progettata per collegare tra loro università e istituti di ricerca in modo da permettere l'accesso alle risorse software e di scambiare messaggi per svolgere una ricerca collaborativa. Tutti sappiamo come questa rete si è sviluppata, quasi come un organismo vivente, alimentandosi dei frutti delle più recenti tecnologie per raggiungere il grande pubblico in quasi tutti gli angoli del pianeta. Arricchita dalle strutture software *World Wide Web*, messe a punto dal CERN in Europa, essa è diventata una biblioteca internazionale di risorse, per certi versi ancora caotica, ma sempre più organizzata e impiegata man mano che i "motori di ricerca" migliorano e incrementano la loro potenza e intelligenza.

E allora, quali insegnamenti possiamo trarre per il futuro da questa storia così affascinante e complessa? A mio parere questa domanda va riformulata in due parti: anzitutto ci si può chiedere quali siano oggi le sfide e le opportunità sulla base dell'attuale stato della tecnologia delle telecomunicazioni numeriche; in secondo luogo quali siano le strutture economiche politiche e sociali che potenziano o che vincolano le nostre capacità collettive di raggiungere questi obiettivi. Alla prima domanda si può rispondere direttamente mentre alla seconda occorre dare una risposta più articolata.

La mia prima risposta può essere riassunta in due esigenze percettibili, per le quali esiste sicuramente un mercato: "dati, conoscenze e passatempi interattivi a portata di mouse" e "comunicazioni individuali ovunque e in qualunque momento". Anche se apparentemente semplici e ovvie, le due definizioni comportano implicazioni sociali complesse. In relazione alla prima, è oggi in atto in USA un conflitto aspro e ancora irrisolto tra i costruttori di televisori e i gestori del servizio di radiodiffusione da una parte, e l'industria dei Personal Computer dall'altra, per quanto concerne gli standard della televisione numerica ad alta definizione (come vedremo in seguito). Al di là del vantaggio competitivo al quale le controparti ambiscono, esistono gravi implicazioni. Mentre è estrema-



THE WHITE HOUSE

WASHINGTON

April 24, 1990

I am pleased to extend warm greetings and congratulations to all those gathered for the 16th Marconi International Fellowship Awards. I also send my special greetings and congratulations to this year's award winner, Dr. Andrew J. Viterbi.

Thanks to the extraordinary breakthroughs of scientists such as Dr. Viterbi, people on opposite sides of the world can be linked together through clear and easy-to-use communications devices. Our Nation is fortunate to have such talented and dedicated scientists whose contributions not only provide marvelous inventions for society but also play a vital role in helping the United States maintain its competitive position in world markets. Clearly, science and technology are inextricably intertwined with our economy, and our Nation must continue to support efforts to foster creative and effective development of new technologies. The Marconi International Fellowship Award is one way of encouraging our scientists, and I applaud the efforts of this program. I also commend Dr. Viterbi for his excellent scientific achievements and for his many contributions to communications.

Barbara joins me in sending best wishes for a memorable and enjoyable award ceremony. God bless you.

George Bush

George Bush, Presidente degli Stati Uniti, si congratula con Andrew J. Viterbi per il premio "16th Marconi International Fellowship Award" (1990).

with both machines and one another, or does it prefer to remain passive to what it receives? And even beyond this, is humankind generally inquisitive and eager to learn or unimaginative and content to watch "game shows"?

Turning now to the demand for "Anywhere, Anytime" digital communication, the words imply mobility and hence wireless connectivity, for which the market is apparent in every street, restaurant and

mente probabile che l'informazione e l'intrattenimento in qualche modo si combineranno, diventando entrambi interattivi, si pone il problema su come reagirà il grande pubblico. Vorrà veramente partecipare attivamente confrontandosi con macchine e interlocutori, oppure preferirà rimanere passivo di fronte a ciò che riceve? E, ancora, il genere umano è effettivamente curioso e desideroso di apprendere oppure è poco intraprendente ed è contento di rimanere spetta-

meeting hall in the industrialized world, making the cellular industry the fastest growing market segment in many countries. But the words also have a much deeper implication. While we take ubiquitous voice communication for granted, for most inhabitants of underdeveloped nations this is an unknown luxury. It is estimated that at least half the world's population has never placed nor received a telephone call, and the most highly populated Asian nations have approximately one telephone line per hundred inhabitants, as compared to the world average of one line per ten inhabitants, and per two inhabitants in developed countries. Mobility or even ubiquity is thus not an issue in these nations, yet to bring them merely to the world average, which is the goal of China and India for the next decade, will require installation of several hundreds of millions of lines. Practically, this goal can only be met by wireless telephony and only efficient digital technologies can provide such capacities within reasonable spectrum allocations.

tore delle trasmissioni a quiz?

Passando alla comunicazione numerica ovunque e in qualunque momento, questo implica mobilità e quindi connettività via radio. Per questo tipo di comunicazioni nel mondo industrializzato esiste un mercato visibile in ogni strada, ristorante e luogo di incontro, che ha fatto dell'industria dei cellulari il segmento di mercato con la maggior crescita in molti Paesi. Si presenta tuttavia una implicazione più profonda. Mentre per noi l'ubiquità della comunicazione fonica è data per acquisita, per la maggior parte degli abitanti dei Paesi in via di sviluppo è un lusso sconosciuto. È stato stimato che almeno metà della popolazione mondiale non ha mai fatto o ricevuto una telefonata e che le nazioni asiatiche più densamente popolate dispongono di circa una linea telefonica per cento abitanti, rispetto ad una media mondiale di una linea per dieci abitanti, e di una linea ogni due abitanti nei Paesi avanzati.

La mobilità o addirittura l'ubiquità non rivestono

L'Aula Magna "Pietro Gismondi" dell'Università di Roma Tor Vergata durante la Lectio Doctoralis. In prima fila (quarto da sinistra) il Preside della Facoltà di Ingegneria dell'Università di Parma Professor Giancarlo Prati assieme (da sinistra) ai Rappresentanti delle Facoltà di Economia, Lettere e Filosofia, Medicina e Chirurgia, Scienze Matematiche Fisiche e Naturali di Tor Vergata.

Let me turn finally to the economic, political and societal structures which may facilitate or impede these efforts. There appear to be some trends and approaches which are common among all continents and others in which Europe, Asia and North America differ widely. Among the common elements, two stand out:

a) *Competition and divestiture in basic services*. A number of U.S. Federal Court decisions over the past three decades have empowered competition in what was once the virtual monopoly of the Bell Telephone System, culminating in its divestiture of all local telephone service to regional operating companies. Though in a variety of different ways and at widely differing paces, this trend has spread worldwide, particularly in wireless telephony. In the U.S. there are several cities that already have four or more cellular service providers and in major European and Asian cities

quindi alcuna rilevanza in questi Paesi; piuttosto per portarli almeno alla media mondiale, obiettivo di Cina e India per il prossimo decennio, richiederà l'installazione di molte centinaia di milioni di linee. Praticamente questo obiettivo può essere raggiunto soltanto con la telefonia senza fili; e solo efficienti tecnologie numeriche potranno rispondere a queste esigenze con ragionevoli attribuzioni dello spettro radio.

Per finire vorrei soffermarmi brevemente sulle strutture economiche, politiche e sociali che possono agevolare od ostacolare questi sforzi. Vi sono alcune evidenti tendenze e orientamenti comuni a tutti i continenti e altri che variano notevolmente tra Europa, Asia e Nord America. Tra gli elementi comuni possono essere enucleati due principali:

a) *Concorrenza e fine della gestione in monopolio dei servizi di base*. Numerose sentenze del tribunale federale degli Stati Uniti nel corso degli ultimi

there are two or more. All this widens the market and promotes competition among manufacturers as well as service providers, and is generally accepted as healthy for all parties.

b) *Globalization.* This is happening at an accelerating pace in a variety of sectors and manners. First, technology is global; corporations on all industrialized continents have common knowledge, tools and techniques. In fact, virtually every major electronic manufacturer has solid-state foundries, or the services of such, on every industrialized continent. Similarly, markets are global. While the European Union is in the lead, having virtually removed tariffs among its member states, other continents are following and it appears that the World Trade Organization will slowly lead to a worldwide lowering of barriers. But even without this, major manufacturers worldwide produce most of their product locally where the goods are sold, thus avoiding most of the tariffs.

Similarly, most major telecommunication service providers are forming global alliances in both long-distance and local wireless markets, with new cross-investments between Europe and North and South America being completed at an ever increasing pace.

Now let us turn to the differences.

Viterbi firma il registro di ammissione alla U.S. National Academy of Sciences. Di fianco Peter H. Raven, Segretario dell'Accademia e, sullo sfondo, il Presidente Bruce Alberts (26 aprile 1997).

1) Investment in Innovation

This refers primarily to research and development of new technologies as well as education of new generations of scientists and engineers. Here the U.S. is currently at a marked disadvantage relative to Europe and Japan. For it has lost the impetus of its former industrial policy and it is groping to modify and revive it. This is one of the less well known consequences of the end of the Cold War. With diminished justification and need for defense expenditures, the Defense Department R&D budget has also been

tre decenni hanno prodotto l'aumento della concorrenza in quello che in precedenza era un monopolio virtuale del Bell Telephone System, che è culminato nel totale trasferimento del servizio di telefonia locale a società di gestione regionale. Anche se con diverse modalità e con tempi largamente diversi, questa tendenza si è diffusa in tutto il mondo, particolarmente nel settore della radiotelefonia. Diverse città degli Stati Uniti d'America hanno già quattro o più gestori di reti di telefonia cellulare mentre nelle principali città europee e asiatiche sono presenti almeno due operatori. Tutto ciò allarga il mercato e promuove la concorrenza sia tra i costruttori che tra i gestori delle reti ed è generalmente ritenuto salutare per tutti.

b) *Globalizzazione.* Questo fenomeno si sta verificando a velocità crescente in una varietà di settori e con modalità differenti. Innanzitutto, la tecnologia è globale: le grandi imprese di tutti i continenti industrializzati hanno conoscenze, mezzi e tecniche comuni. Infatti, in pratica tutti i principali fabbricanti del settore elettronico dispongono di impianti per la produzione dei circuiti a stato solido, o comunque hanno accesso ai relativi servizi in ogni continente industrializzato. In modo analogo i mercati sono anch'essi globali. Mentre l'Unione Europea è al primo posto, avendo abolito quasi del tutto i dazi tra gli Stati membri, altri continenti seguono a ruota e l'Organizzazione mondiale del commercio (WTO) sembra voler portare lentamente ad un abbassamento generalizzato delle barriere doganali. Comunque, anche senza questi cambiamenti, i principali produttori in tutto il mondo fabbricano gran parte dei loro prodotti là dove questi sono commercializzati evitando quindi la maggior parte dei dazi.

Analogamente, la maggioranza dei gestori di servizi di telecomunicazioni stanno creando alleanze globali sui mercati sia sui sistemi radio locali sia in quelli a lunga distanza, predisponendo nuovi investimenti tra l'Europa e le Americhe a ritmo crescente.

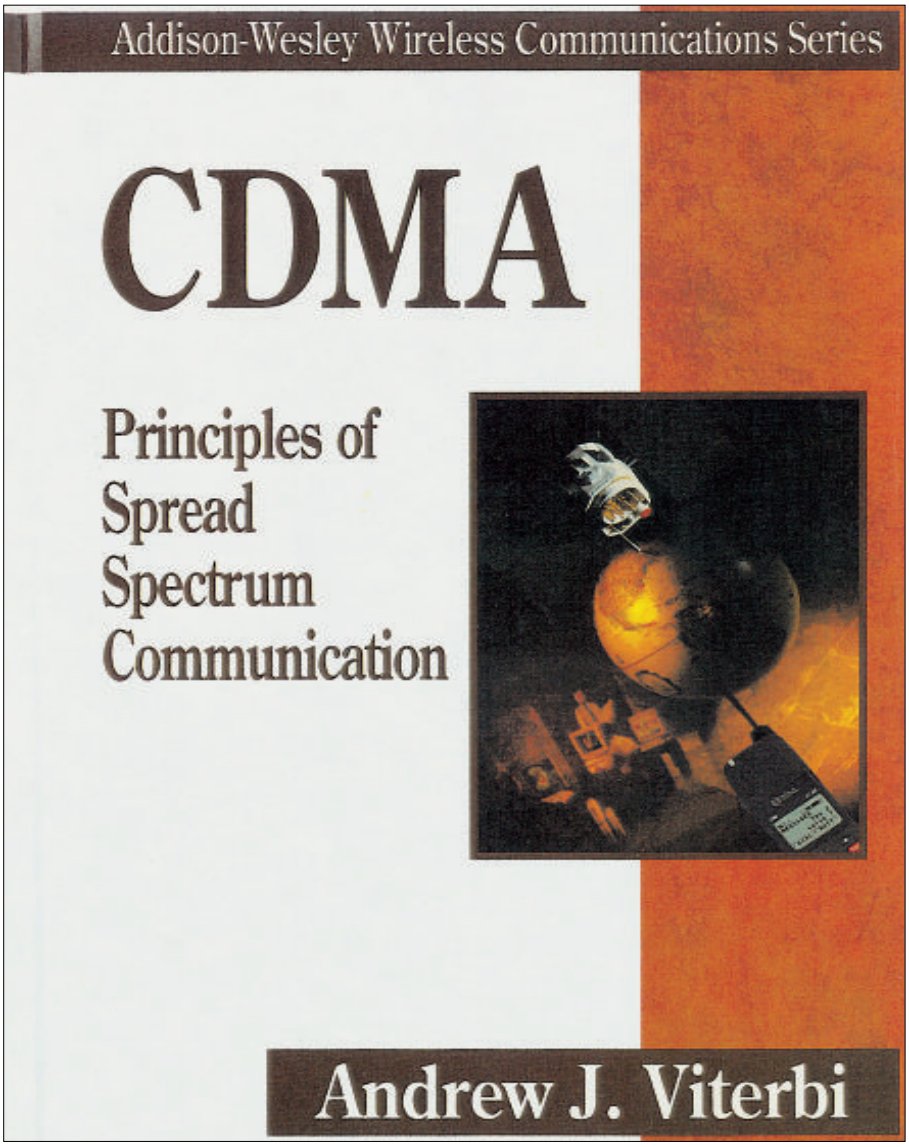
Passiamo ora alle differenze.

1) Investire nell'innovazione

Questo punto si riferisce principalmente alla ricerca e allo sviluppo di nuove tecnologie e all'istruzione di nuove generazioni di ricercatori e di tecnici. In questo campo gli Stati Uniti sono ora in una posizione di notevole svantaggio rispetto all'Europa e al Giappone, in quanto l'impeto della precedente politica industriale stenta a riprendersi nonostante gli

curtailed. This has had an adverse effect on advanced scientific and technological education, since some of those research funds went to the training of new researchers.

Europe, on the other hand, both by individual major countries and through the EU has provided significant and ever increasing research and development support to industry and universities through programs like ESPRIT, RACE and ACTS. Japan's MITI and Korea's ETRI are governmental organizations fostering and supporting R&D in a variety of different ways.



Il libro di Andrew J. Viterbi sulle tecniche di accesso a divisione di codice (CDMA).

Another vehicle for investment in innovation in the U.S., sometimes envied by other nations, is the venture capitalist community, which invests in start-ups in return for a significant equity position. While it is true that some of the pioneering Silicon Valley companies, previously referred to *(see page 72)*, needed such financing to get started, this funding is not for visionary research which often takes several years to reach development maturity. The venture

sforzi per modificarla e rinvigorirla. Questa è una delle conseguenze meno note della fine della guerra fredda. Con la diminuita giustificazione ed esigenza di spese per la difesa, il bilancio per ricerca e sviluppo del Ministero della Difesa ha subito pesanti tagli che a loro volta hanno avuto notevoli effetti negativi sull'istruzione avanzata in campo scientifico e tecnologico, poiché parte dei fondi per la ricerca servivano per l'addestramento di nuovi ricercatori.

In Europa, invece, è andata diversamente: sia i singoli Paesi più importanti sia l'Unione Europea hanno appoggiato in modo significativo e crescente gli sforzi di ricerca e sviluppo dell'industria e delle università tramite il finanziamento di programmi come ESPRIT, RACE e ACTS. Il MITI in Giappone e la ETRI coreana sono Enti governativi di promozione e di supporto delle attività di ricerca e sviluppo con modalità diverse.

Un altro veicolo per gli investimenti nell'innovazione negli Stati Uniti, talvolta invidiati dagli altri Paesi, è rappresentato dalla comunità di investitori di "venture capital" che puntano su nuove iniziative in cambio di un corrispettivo rilevante in pacchetti azionari. Anche se è vero che alcune delle aziende pioniere della Silicon Valley citate in precedenza *(vedi pagina 72)* avevano bisogno di questi finanziamenti per avviare l'attività, questi fondi non sono indirizzati alla ricerca di lungo periodo che in molti casi richiede anni prima di portare a sviluppi concreti. Il venture capital è raramente impegnato per più di uno o due anni. Contemporaneamente, con la fine dei monopoli e l'aumento della concorrenza, i grandi istituti di ricerca come i Laboratori Bell hanno notevolmente ridotto il livello delle spese per la ricerca di base a lungo termine.

2) Più regole o meno regole

Nella politica delle telecomunicazioni i continenti divergono notevolmente. Almeno nell'ultimo decennio, la *FCC (Federal Communication Commission)* degli Stati Uniti ha evitato di definire qualunque tipo di standard limitandosi ad attribuire porzioni di spettro radio per i vari servizi. Invece, ha spronato sia i gestori di servizi che i fabbricanti a definire standard appropriati tramite le associazioni di enti industriali. Non soltanto questo è valido particolarmente per gli standard dei sistemi cellulari numerici; ma persino

capitalist vistas are rarely more than a year or two into the future. At the same time, with divestiture and increased competition, corporate research institutes like Bell Laboratories have considerably reduced their level of expenditures on fundamental long-term research.

2) Regulation or Its Avoidance

Here the continents diverge seriously in telecommunication policy. The U.S. *FCC (Federal Communication Commission)* for at least the past decade has totally avoided setting standards and has limited its role to spectrum allocation for various services. It has instead encouraged industry, both service providers and manufacturers, to establish standards through industry associations. Not only has this been clearly the case for digital cellular, but even for digital television standards, which began with active FCC involvement, at the eleventh hour its approval was compromised, due to the conflicting views of the television and personal computer industries, as noted prevlously.

European administrations, in contrast, control standards as well as spectrum. Asian governments do as well, although of late Japan has tended to be more pluralistic. But really, with or without govemment control, what matters is the standard itself and how it is established, which we treat next.

3) Standardization

Standards bodies exist on all continents and in all industrialized nations and regions. For telecommunication, in Europe it is the *ETSI (European Telecommunication Standards Institute)*, in Japan it is the *ARIB (Association of Radio Industries and Businesses)*, and in North America the *TIA* (*Telecommunications Industry Association)*.

Surprisingly they operate in similar ways, although some are more receptive than others to new initiatives.

The major difference, however, seems to be on emphasis. In all cases there is an inherent and natural tension between three strong forces: market requirements, technology and policy. In an ideal world the three should be mutually supportive, but hardly ever is this utopian fantasy achieved. The contrast in the three major industrialized regions is summarized by the relative weight given to these three factors. In the U.S., it is marketplace first, technology second and policy non-existent, as already noted. In Europe, perhaps the inverse order holds. In Japan it is arguably market first, policy a close second and technology third. We all have our opinions on the wisdom of each approach, but like our personal views on finances, religion and politics, it is best at times to keep them to ourselves.

per quelli della televisione numerica, i cui inizi hanno visto l'attiva partecipazione della FCC, non è stato possibile raggiungere un compromesso in extremis per le opinioni conflittuali delle industrie della televisione e dei personal computer come indicato in precedenza.

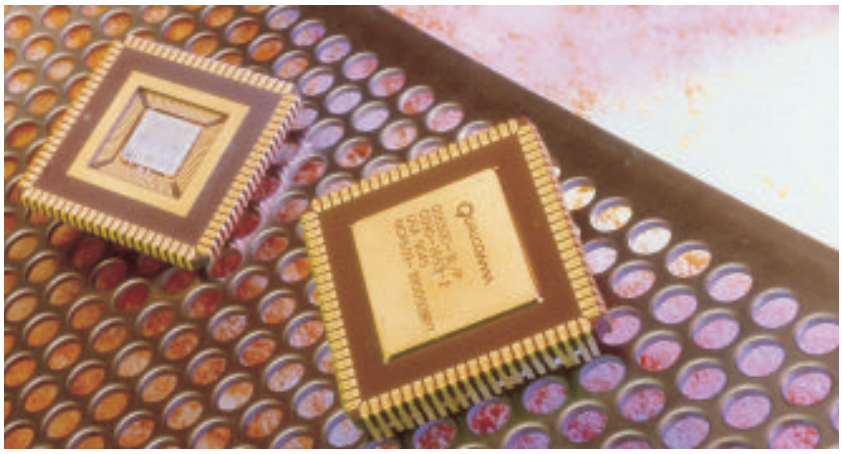
Le amministrazioni europee controllano invece sia gli standard che lo spettro radio. I governi asiatici fanno lo stesso, anche se ultimamente il Giappone ha optato per un maggiore pluralismo. Tuttavia, con o senza controllo governativo, ciò che più conta è lo standard di per sé e come esso è stabilito, argomento del prossimo punto.

3) Normalizzazione

Gli enti di standardizzazione sono presenti su tutti i continenti e in tutti i Paesi e le aree industrializzate. Per le telecomunicazioni, in Europa esiste l'*ETSI* (*European Telecommunication Standards Institute)*, in Giappone vi è l'*ARIB* (*Association of Radio Industries and Businesses)* e in Nord America la *TIA (Telecommunications Industry Association)*.

È sorprendente come questi Enti operino con modalità simili anche se alcuni sono più sensibili di altri alle nuove iniziative.

Le differenze maggiori sembrano tuttavia essere sulle priorità. È sempre presente una conflittualità fisiologica tra tre potenti forze in gioco: esigenze di mercato, tecnologia e politica. In un mondo perfetto le tre componenti dovrebbero complementarsi a vicenda, cosa che purtroppo rimane troppo spesso un'utopia. I contrasti nell'ambito delle tre maggiori regioni industrializzate possono essere riassunti dal peso relativo dato ai tre fattori. Negli Stati Uniti vi è in primo piano il mercato, seguito dalla tecnologia mentre le politiche sono come già detto inesistenti. In Europa, è forse il contrario. In Giappone, si potrebbe sostenere che venga prima il mercato seguito da vicino dalle politiche, mentre la tecnologia occupa il terzo posto. Ognuno di noi ha le sue opinioni sulla validità di ciascun approccio, ma come per le convinzioni personali sulla finanza, sulla religione e sulla politica, talvolta queste è meglio tenerle per sé.



Un recente circuito integrato della Qualcomm, con decodificatore Viterbi ad alta velocità.

Conferimento laurea Honoris Causa

Placet del Senato Accademico

Il Rettore Alessandro Finazzi Agrò consegna la pergamena a Viterbi.

Universitatis Rector: Iudices, luculenta dissertatione audita, vestram sententiam patefacite!

Facultatis Decanus: Facultatis nostrae magistri, magna ornatissimi doctrina, placetne dominum Andream Viterbi, terra ab illa quae ex Americi nomine America vocatur Romam petitum, doctorem Honoris Causa artis eminus electrica vi nuntiandi fieri?

Professorum Collegium: Placet.

Facultatis Decanus: Magnifice Rector, Magistri omnes artium uno consilio, uno animo, una voluntate putant doctissimum dominum Andream Viterbi Honoris Causa academicis lauris praecingendum esse. Te ergo una voce rogamus ut ille tua manu hac magna dignitate decoretur.

Il Preside Franco Maceri si congratula con Viterbi.

ROMANA STVDIORVM VNIVERSITAS
"AD TVRRIM VIRGATAM"

NOS ALEXANDER FINAZZI AGRÒ
VNIVERSITATIS RECTOR

VNO ANIMO EXOPTANTIBVS AEDILIS ET MACHINALIS ARTIS PROFESSORIBVS
VT ACADEMICA DIGNITATE H.C. DECORETVR ANDREAS JACOBVS VITERBI
QVIPPE QVI NOVO ALGORISMO EXCOGITATO
EFFECERIT VT PER NVMERORVM CONTINVATIONEM SERIEMQVE
QVOVIS DEFERANTVR RERVM COGNITIONES ET NOTITIAE
NECNON PLANE ATQVE PERSPICVE
INTERVALLO INTERIECTO QVAMVIS LONGO PERCIPIANTVR
IDEMQVE ETIAM ATQVE ETIAM VEL DISSERENDO VEL EXPERIENDO
QVID VTILITATIS AFFERANT CELLVLAE ELECTRICOMAGNETICIS VNDIS NIXAE
NOTIONVM COMMERCIIS PER TELEPHONIVM INSTITVTIS EXPLANAVERIT
DEINDE NOVAS RATIONES NVNTIANDI
OPE MACHINAE IN INANI SITAE PROTVLERIT
ITA VT NVNTII EMISSI ATQVE REVERSI EA VIA OMNIBVS GENTIBVS
PRAESTO SINT ET MVLTIMODIS IN VSVM CONVERTANTVR

HIS DE CAVSIS

ANDREAM JACOBVM VITERBI

OMNI LAVDE DIGNVM OB OPERA PROVECTA
DOCTOREM HONORIS CAVSA IN ARTE EMINVS ELECTRICA VI NVNTIANDI
DECLARAMVS ET RENVNTIAMVS
EIQVE OMNIA HVIVS GRADVS PRIVILEGIA HONORES IVRA CONFERIMVS
IN CVIVS REI FIDEM HOC DIPLOMA
VNIVERSITATIS SIGILLO MVNITVM
EI TRADIMVS

DATVM ROMAE DIE XIII MENS. MAI. A.D. MCMLXXXXVII

STVDIORVM VNIVERSITATIS RECTOR

SECRETARIVS GENERALIS

FACVLTATIS DECANVS

Il diploma della laurea Honoris Causa ad Andrew J. Viterbi.

Il Notiziario Tecnico Telecom Italia, ritenendo di fare cosa gradita ai lettori, ha riprodotto due articoli di rilievo pubblicati dal professor Andrew J. Viterbi. La Redazione della Rivista ringrazia la Direzione di IEEE Transactions per aver permesso di riprodurre i due testi.

IEEE TRANSACTIONS ON

INFORMATION THEORY

APRIL 1967 VOLUME IT-13 NUMBER 2

Published Quarterly A Journal Devoted to the Theoretical Experimental Aspects of Information Transmission, Processing, and Utilization

Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm

ANDREW J. VITERBI, SENIOR MEMBER, IEEE

***Abstract*—The probability of error in decoding an optimal convolutional code transmitted over a memoryless channel is bounded from above and below as a function of the constraint length of the code. For all but pathological channels the bounds are asymptotically (exponentially) tight for rates above R_0, the computational cutoff rate of sequential decoding. As a function of constraint length the performance of optimal convolutional codes is shown to be superior to that of block codes of the same length, the relative improvement increasing with rate. The upper bound is obtained for a specific probabilistic nonsequential decoding algorithm which is shown to be asymptotically optimum for rates above R_0 and whose performance bears certain similarities to that of sequential decoding algorithms.**

Manuscript received May 20, 1966; revised November 14, 1966. The research for this work was sponsored by Applied Mathematics Division, Office of Aerospace Research, U. S. Air Force, Grant AFOSR-700-65.

The author is with the Department of Engineering, University of California, Los Angeles, Calif.

I. Summary of Results

SINCE Elias[1] first proposed the use of convolutional (tree) codes for the discrete memoryless channel, it has been conjectured that the performance of this class of codes is potentially superior to that of block codes of the same length. The first quantitative verification of this conjecture was due to Yudkin[2] who obtained

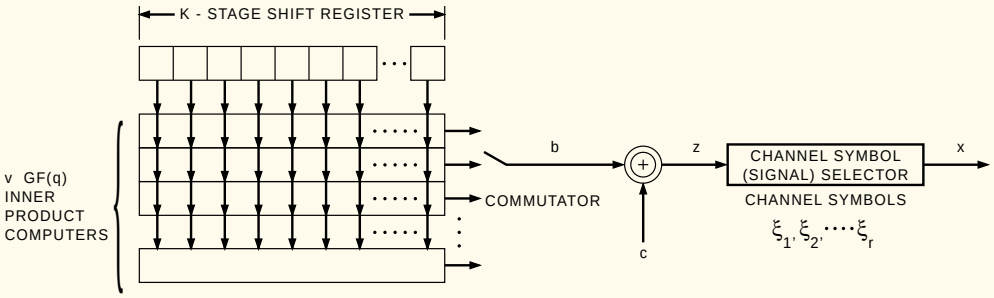


Fig. 1. Encoder for q-ary convolutional (tree) code.

an upper bound on the error probability of an optimal convolutional code as a function of its constraint length, which is achieved when the Fano sequential decoding algorithm[3] is employed.

In this paper, we obtain a lower bound on the error probability of an optimal convolutional code independent of the decoding algorithm, which for all but pathological channels is asymptotically (exponentially) equal to the upper bound for rates above R_0, the computational cutoff rate of sequential decoding. Also, a new probabilistic nonsequential decoding algorithm is described, which exhibits and exploits a fundamental property of convolutional codes. An upper bound on error probability utilizing this decoding algorithm is derived by random coding arguments, which coincides with the upper bound of Yudkin.[2] In the limit of very noisy channels, upper and lower bounds are shown to coincide asymptotically (exponentially) for all rates, and the negative exponent of the error probability, also known as the reliability, is shown to be

$$\lim_{N\to\infty} \frac{1}{N} \ln (1/P_E) = \begin{cases} C/2 & 0 \leq R \leq C/2 \\ C - R & C/2 \leq R < C \end{cases}$$

where N is the code constraint length (in channel symbols), R is the transmission rate and C is channel capacity. This represents a considerable improvement over block codes for the same channels. Also, it is shown that in general in the neighborhood of capacity, the negative exponent is linear in $(C - R)$ rather than quadratic, as is the case for block codes.

Finally, a semisequential modification of the decoding algorithm is described which has several of the basic properties of sequential decoding methods.[3],[4]

II. Description and Properties of the Encoder

The message to be transmitted is assumed to be encoded into the data sequence **a** whose components are elements of the finite field of q elements, $GF(q)$, where q is a prime or a power of a prime. All messages are assumed equally likely; hence all sequences **a** of a fixed number of symbols are equally probable. The encoder consists of a K-stage shift register, v inner-product computers, and an adder, all operating over $GF(q)$, together with a channel symbol selector connected as shown in Fig. 1. After each q-ary symbol of the sequence is shifted into the shift register, the uth computer ($u = 1, 2, \cdots v$) forms the inner product of the vector in the shift register, which is a subsequence of **a**, with some fixed K-dimensional vector $\mathbf{g}_u$, whose components are also elements of $GF(q)$. The result is a matrix multiplication of the K symbol subsequence of **a** (as a row vector) with a Kxv matrix G (whose uth column is $\mathbf{g}_u$) to produce v symbols of the sequence **b**. This is added to v symbols of a previously stored (or generated) q-ary sequence **c**, whose total length is $(L + K - 1)v$ symbols. The v symbol subsequence of **z** thus generated can be any one of q^v v-component vectors. By properly selecting the matrix G and subsequence of **c** [or by selecting them at random with uniform probability from among the ensemble of all q^{Kv} matrices and q^v vectors with components in $GF(q)$], all possible v symbol subsequences of **z** can be made to occur with equal probability. Finally the channel symbol selection (or signal selection in the case of continuous channels) consists of a mapping of each q-ary symbol of **z** onto an r-ary channel symbol x_i of the channel input sequence **x** (where $r \leq q$), as follows: let n_1 of the q-ary symbols be mapped into ξ_1, n_2 into ξ_2, etc., such that

$$\sum_{i=1}^{r} n_i = q.$$

Thus if each symbol of **z** is with uniform probability any element of $GF(q)$, the probability distribution of the jth channel input symbol x_j is

$$P(x_j = \xi_i) = \frac{n_i}{q} \qquad (i = 1, 2, \cdots r) \quad \text{for all} \quad j$$

and by proper choice of q and r any rational channel input distribution can be attained. Furthermore, since one q-ary data symbol thus produces v channel symbols, the transmission rate of the system is

$$R = \frac{\ln q}{v} \frac{\text{nats}}{\text{channel symbol}} \tag{1}$$

and thus, by proper choice of q (which must be a prime or the power of a prime) and v, any rate can be closely approximated.

We note also that the encoder thus produces a tree code with q branches, each containing v channel symbols, emanating from each branching node since for every

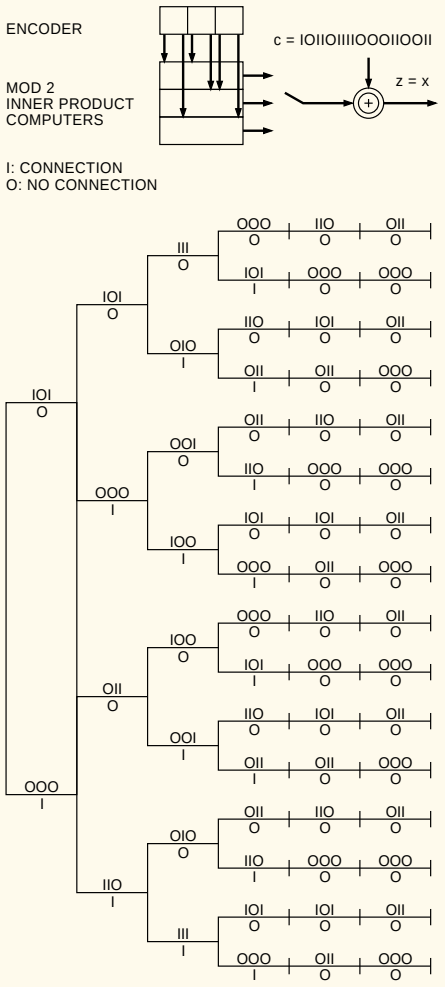


Fig. 2. Tree code for $q = 2$, $v = 3$, $r = 2$, $L = 4$, $K = 3$.

shift of the register a potentially different set of v channel symbols is generated for each of the q possible values of the data symbol. An example is shown in Fig. 2 for $q = 2$, $v = 3$, $r = 2$, $K = 3$. The data symbol a_i is indicated below each branch while the channel symbols x_i are indicated above each branch.

The procedure continues until L data symbols are fed into the shift register followed by a sequence of $K - 1$ zeros. L is known as the (*branch*) *tree length*, and $N = Kv$ as the (*symbol*) *constraint length* of the code. The overall encoding algorithm thus produces a tree code with L branching levels. All branches contain v channel symbols except for the q^L final branches which contain $N = Kv$ channel symbols. The example of Fig. 2 shows such a tree code for $L = 4$ and $K = 3$.

A basic property of the convolutional code thus generated by the K-stage shift register is the following.

A) Two divergent paths of the tree code will converge (i.e., produce the same channel symbols) after the data symbols corresponding to the two paths have been identical for K consecutive branches. Two paths are said to be *totally distinct* over any sequence of branches for which this event does not occur.

We now proceed to derive the lower bound on error probability for an optimal convolutional code using property A) and lower bound results for optimal block codes.

III. The Lower Bound

Suppose a magic genie informs the decoder as to the exact state of each branch data symbol a_i for all branches i ($i = 1, 2, \cdots L + K - 1$) except for the m consecutive branches $j + 1, j + 2, \cdots j + m (0 \leq j \leq L - m)$. Thus to decode the tree the decoder must decide upon which of the q^m possible m-symbol q-ary data sequences corresponding to these m branches actually occurred, or equivalently he must decide among the corresponding q^m alternate paths through the tree. To do this he has available the $(L + K - 1)v$ symbol received tree sequence $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots \mathbf{y}_{L+K-1})$ where $\mathbf{y}_i$ is the received symbol sequence for the ith branch. Actually since the a_i are known for all $i \leq j$, he needs only examine $\mathbf{y}_i$ for $i \geq j + 1$. Furthermore, the q^m alternate paths in question, which diverge at the $(j + 1)$th branch must converge again at the $(j + m + K)$th branch, for since all the corresponding branch data symbols a_i are identical for $i \geq j + m + 1$, by the $(j + m + K)$th branch the data symbols in the shift register will be identical for all paths in question. Thus the q^m paths are totally distinct over at most $m + K - 1$ branches. Now letting

$$\mu = \frac{m}{K} \tag{2}$$

and having denoted the constraint length in channel symbols by

$$N = Kv \tag{3}$$

we obtain from (1), (2), and (3)

$$m \ln q = \mu NR. \tag{4}$$

The optimal decoder for paths which are a priori equally likely must compute the $q^m = e^{\mu NR}$ likelihood functions $p(\mathbf{y} \mid \mathbf{a})$, where $\mathbf{a} = (a_{j+1}, \cdots a_{j+m})$ is an m-component q-ary vector which specifies the path, and $\mathbf{y} = (\mathbf{y}_{j+1}, \cdots \mathbf{y}_{j+m+K-1})$ is an $(m + K - 1)v = (\mu + 1)N - v$ component vector, and select the path corresponding to the greatest. The resulting error probability is lower bounded by the lower bound[5]-[7] for the best block code with $e^{\mu NR}$ words of length $(\mu + 1)N - v$ channel symbols transmitted over a memoryless channel with discrete input space:

$$P_E(\mu, N, R) > \exp\{-N(\mu + 1)[E_L(R, \mu) + o(\mu N)]\} \tag{5}$$

where

$$o(\mu N) \to 0 \quad \text{linearly} \quad \mu N \to \infty$$
$$E_L(R, \mu) = \underset{0 \leq \rho \leq \infty}{\text{l.u.b.}} \left[\hat{E}_0(\rho) - \rho \frac{\mu}{\mu + 1} R\right] \tag{6}$$

and $\hat{E}_0(\rho)$ is the concave hull of the function

$$E_0(\rho) = \max_{p(x)} \{-\ln \sum_Y [\sum_X p(x)p(y \mid x)^{1/1+\rho}]^{1+\rho}\} \quad (7)$$

where X and Y are the channel input and output spaces, respectively, $p(y \mid x)$ is the channel transition probability distribution, and $p(x)$ is an arbitrary probability distribution on the input space. Furthermore, the function $E_0(\rho)$ has the following basic properties which are proved in Gallager:[5]

a) $E_0(0) = 0$ and $E_0(\rho) > 0$ for all $\rho > 0$,

b) $E_0'(\rho) > 0$ for all finite ρ, and $\lim_{\rho \to 0} E_0'(\rho) = C$ which is the channel capacity.

For most channels of interest $E_0(\rho)$ is itself a concave function. When this is not the case the channel is said to be pathological.[5]

This bound, known as the sphere-packing bound, is the tightest exponential bound for high rates. For low rates a tighter bound, which has been recently derived,[7] is considered below. $E_L(R, \mu)$ can be obtained by solving the parametric equations

$$E_L(R, \mu) = \hat{E}_0(\rho) - \rho \hat{E}_0'(\rho) \quad (8a)$$

$$R = \frac{\mu + 1}{\mu} \hat{E}_0'(\rho). \quad (8b)$$

But $\mu = m/K$ can be any multiple of $1/K$ up to L/K, since m cannot exceed L. Hence, since no particular demands can be made on the magic genie,

$$P_E(N, R) \geq \max_{(1/K) \leq \mu \leq (L/K)} P_E(\mu, N, R)$$

$$> \exp \{-N \min_{(1/K) \leq \mu \leq (L/K)} (\mu + 1)[E_L(R, \mu) + o(\mu N)]\} \quad (9)$$

corresponding to the least obliging genie for the particular R.

Thus we seek the lower envelope

$$E_L(R) = \min_{(1/K) \leq \mu \leq (L/K)} (\mu + 1)E_L(R, \mu). \quad (10)$$

It follows from (6) and (7) and property b) that

$$\lim_{\mu \to 0} (\mu + 1)E_L(R, \mu) = \text{l.u.b.}_{0 \leq \rho \leq \infty} \hat{E}_0(\rho) = \hat{E}_0(\infty)$$

$$\lim_{\mu \to \infty} (\mu + 1)E_L(R, \mu) = \infty \quad \text{for} \quad R < C.$$

The family of functions $(\mu + 1)E_L(R, \mu)$ is sketched in Fig. 3. To find the lower envelope we must minimize $E_L(R, \mu)$ over the set of possible μ for each R. For the purposes of the lower bound we shall let L/K be as large as required for the minimization. First, let us minimize over all positive real μ and then restrict μ to be a multiple of $1/K$. Thus from (8a) we have

$$\frac{\partial[(\mu + 1)E_L(R, \mu)]}{\partial \mu}$$

$$= \hat{E}_0(\rho) - \rho \hat{E}_0'(\rho) + (\mu + 1)[-\rho \hat{E}_0''(\rho)] \frac{\partial \rho}{\partial \mu} \quad (11)$$

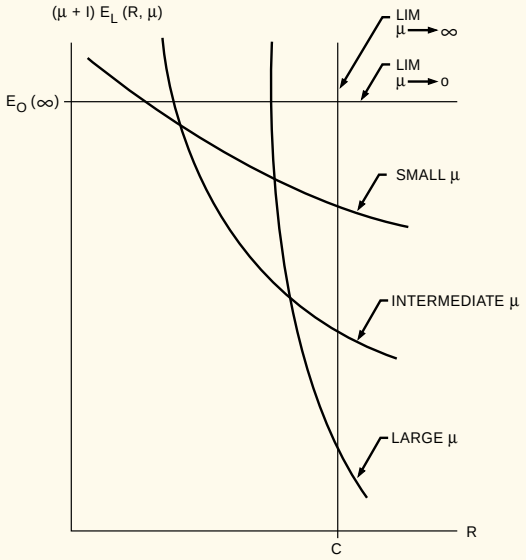


Fig. 3. Family of funcitions $(\mu + 1)\, E_L\, (R, \mu)$.

while from (8b) we have

$$\hat{E}_0''(\rho) \frac{\partial \rho}{\partial \mu} = \frac{R}{(\mu + 1)^2}. \quad (12)$$

Combining (11) and (12) and setting the former equal to zero, we find that the function has a stationary point at

$$\mu = \frac{\rho R}{\hat{E}_0(\rho) - \rho \hat{E}_0'(\rho)} - 1. \quad (13)$$

Furthermore, differentiating (11) and using (12), we find that

$$\frac{\partial^2[(\mu + 1)E_L(R, \mu)]}{\partial \mu^2} = -\frac{R^2}{(\mu + 1)^3 \hat{E}_0''(\rho)} \geq 0$$

so that (13) corresponds to an absolute minimum. Inserting (13) in (8b) yields

$$R = \frac{\hat{E}_0(\rho)}{\rho} \quad (14)$$

and since $\hat{E}_0(\rho)$ is concave it follows that $R = \hat{E}_0(\rho)/\rho \geq \hat{E}_0'(\rho)$ which implies that the solution (13) for μ is nonnegative. From (8a), (13), and (14) we obtain

$$\min_{0 \leq \mu < \infty} (\mu + 1)E_L(R, \mu) = \rho R = \hat{E}_0(\rho). \quad (15)$$

Now, since μ is restricted to be a multiple of $1/K$, let us consider altering (13) by adding a positive real number δ large enough to make μ an element of this set. In any case $\delta < 1/K$. But changing μ by this amount in (9) alters the exponent by an amount proportional to $N/K = v$, which is a constant parameter of the encoder and hence, normalized by N, is $o(N)$. The rate is also altered by an amount of the order of $1/K$ by this change in μ, but if we adjust for this change by returning R to its original value (14), we again alter P_E by an amount of magnitude $o(N)$. Thus from (9), (10), (14), and (15) we obtain

Theorem 1

The probability of error in decoding an arbitrarily long convolutional code tree of constraint length N (channel symbols) transmitted over a memoryless channel is bounded by

$$P_E > \exp\{-N[E_L(R) + o(N)]\}$$

where

$$E_L(R) = \hat{E}_0(\rho) \qquad (0 \leq \rho < \infty) \tag{16a}$$

and

$$R = \hat{E}_0(\rho)/\rho. \tag{16b}$$

Taking the derivative of (14) we find

$$\frac{\partial R}{\partial \rho} = \frac{\hat{E}_0'(\rho) - \hat{E}_0(\rho)/\rho}{\rho} \leq 0 \quad \text{for all} \quad \rho > 0$$

where we have made use of the fact that $\hat{E}_0(\rho)$ is concave. Also, from property b) we have $\lim_{\rho \to 0} \hat{E}_0(\rho)/\rho = \hat{E}'(0) = C$. Thus we obtain

Corollary 1

The exponent $E_L(R)$ in the lower bound is a positive monotone decreasing continuous function of R for all $0 \leq R < C$.

A graphical construction of the exponent-rate curve from a plot of the function $E_0(\rho)$ is shown in Fig. 4. We defer further consideration of the properties of (16) until after an upper bound is obtained.

A tighter lower bound on error probability for low rates is obtained by replacing the sphere packing bound of (6) by the tighter lower bound for low rates recently obtained by Shannon, Gallager, and Berlekamp.[7] For this bound (6) is replaced by

$$E_L(R, \mu) = E_x - \frac{\tilde{\rho}\mu R}{\mu + 1} \qquad \left(0 \leq R \leq \frac{\mu + 1}{\mu} \hat{E}_0'(\tilde{\rho})\right) \tag{17a}$$

where

$$E_x = \max_{p(x)} \{-\lim_{\rho \to \infty} [\rho \ln \sum_X \sum_X p(x)p(x') \cdot (\sum_Y \sqrt{p(y \mid x)p(y \mid x')})^{1/\rho}]\} = \hat{E}_0(\tilde{\rho}). \tag{17b}$$

The straight line of (17a) is tangent to the curve of (6) at $R = [(\mu + 1)/\mu]\hat{E}_0'(\tilde{\rho})$. Repeating the minimization with respect to μ we find

$$E_L(R) = \min_{\mu} [(\mu + 1)E_x - \tilde{\rho}\mu R]$$

$$= E_x, \qquad 0 \leq R \leq \frac{\hat{E}_0(\tilde{\rho})}{\tilde{\rho}}.$$

Thus, we have

Corollary 2

For low rates a tighter lower bound than that of Theorem 1 is:

$$P_E > \exp\{-N[E_L(R) + o(N)]\}$$

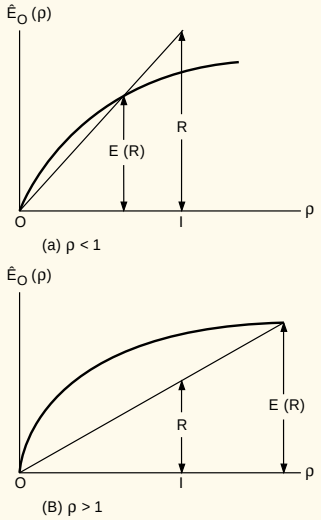


Fig. 4. Graphical construction of $E_L(R)$ from $\hat{E}_O(P)$.

where

$$E_L(R) = E_x, \qquad 0 \leq R \leq \frac{\hat{E}_0(\tilde{\rho})}{\tilde{\rho}}, \tag{18}$$

$\tilde{\rho}$ is the solution to the equation $\hat{E}_0(\tilde{\rho}) = E_x$, and E_x is given by (17b).

IV. A Probabilistic Nonsequential Decoding Algorithm

We now describe a new probabilistic nonsequential decoding algorithm which, as we shall show in the next section, is asymptotically optimum for rates $R > R_0 = E_0(1)$. The algorithm decodes an L-branch tree by performing L repetitions of one basic step. We adopt the convention of denoting each branch of a given path by its data symbol a_i, an element of $GF(q)$. Also, although $GF(q)$ is isomorphic to the integers modulo q only when q is a prime, for the sake of compact notation, we shall use the integer r to denote the rth element of the field.

In *Step 1* the decoder considers all q^K paths for the first K branches (where K is the branch constraint length of the code) and computes all q^K likelihood functions $\prod_{i=1}^{K} p(\mathbf{y}_i \mid a_i)$. The decoder then compares the likelihood function for the q paths:

$$(0, a_2, a_3, \cdots a_K),$$

$$(1, a_2, a_3, \cdots a_K),$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$(q - 1, a_2, a_3, \cdots a_K)$$

for each of the q^{K-1} possible vectors $(a_2, a_3 \cdots a_K)$. It thus performs q^{K-1} comparisons each among q path likelihood functions. Let the path corresponding to the greatest likelihood function in each comparison be denoted the *survivor*. Only the q^{K-1} survivors of as many comparisons are preserved for further consideration; the remaining paths are discarded. Among the q^{K-1} survivors

each of the q^{K-1} vectors $(a_2, a_3, \cdots a_K)$ is represented uniquely, since by the nature of the comparisons no two survivors can agree in this entire subsequence.

Step 2 begins with the computation for each survivor of Step 1 of the likelihood functions of the q branches emanating from the $(K+1)$th branching node and multiplication of each of these functions by the likelihood function for the previous K branches of the particular path. This produces q^K functions for as many paths of length $K+1$ branches, and each of the subsequences $a_2, a_3, \cdots a_{K+1}$ are represented uniquely. Again the q^K functions are compared in groups of q, each comparison being among the set of paths:

$$(\alpha_{11}^{(1)}, 0, a_3, a_4 \cdots a_{K+1})$$
$$(\alpha_{21}^{(1)}, 1, a_3, a_4 \cdots a_{K+1})$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$(\alpha_{q1}^{(1)}, q-1, a_3, a_4 \cdots a_{K+1})$$

where $\alpha_{k1}^{(1)}$ corresponds to the first branch of the survivor of a comparison performed at the first step. Again only the survivors of the set of q^{K-1} comparisons are preserved and the remaining paths are discarded. The algorithm proceeds in this way, at each step increasing the population by a factor of q by considering the set of q branches emanating from each surviving path and then reducing again by this factor by performing a new set of comparisons and excluding all but the survivors.

In particular, at *Step* $j+1$ the decoder performs q^{K-1} sets of comparisons among groups of q paths, which we denote

$$(\alpha_{11}^{(j)}, \alpha_{12}^{(j)}, \cdots \alpha_{1j}^{(j)}, 0, a_{j+2}, a_{j+3}, \cdots a_{j+K}),$$
$$(\alpha_{21}^{(j)}, \alpha_{22}^{(j)}, \cdots \alpha_{2j}^{(j)}, 1, a_{j+2}, a_{j+3}, \cdots a_{j+K}),$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$(\alpha_{q1}^{(j)}, \alpha_{q2}^{(j)} \cdots \alpha_{q,j}^{(j)}, q-1, a_{j+2}, a_{j+3}, \cdots a_{j+K})$$

where the vectors $(\alpha_{k1}^{(j)}, \alpha_{k2}^{(j)}, \cdots \alpha_{kj}^{(j)})$ depend on the outcome of the previous set of comparisons. Again by the nature of the comparisons no two survivors can agree in all of the last $K-1$ branches and there is a one-to-one correspondence between each of the q^{K-1} survivors and the subsequences $(a_{j+2}, \cdots a_{j+K})$.

This procedure is repeated through the $(L-K+1)$th step. Beyond this point branching ceases because only zeros are fed into the shift register. Thus at step $L-K+2$ the decoder compares the likelihood functions for the q paths:

$$(\alpha_{11}^{(L-K+1)}, \alpha_{12}^{(L-K+1)}, \cdots \alpha_{1,L-K+1}^{(L-K+1)}, 0, a_{L-K+3} \cdots a_L, 0),$$
$$(\alpha_{21}^{(L-K+1)}, \alpha_{22}^{(L-K+1)}, \cdots \alpha_{2,L-K+1}^{(L-K+1)}, 1, a_{L-K+3} \cdots a_L, 0),$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$(\alpha_{q1}^{(L-K+1)}, \alpha_{q2}^{(L-K+1)}, \cdots \alpha_{q,L-K+1}^{(L-K+1)}, q-1, a_{L-K+3} \cdots a_L, 0)$$

for each of the q^{K-2} possible vectors $(a_{L-K+3} \cdots a_L)$ resulting in q^{K-2} survivors. Thus, for this and all succeeding steps the population fails to grow since all further branches correspond only to zeros entering the shift register, and it is reduced by a factor of q by the comparisons. Thus, just after the $(L-1)$th step there are only q survivors:

$$(\alpha_{11}^{(L-1)}, \cdots \alpha_{1,L-1}^{(L-1)}, 000 \cdots 0),$$
$$(\alpha_{21}^{(L-1)}, \cdots \alpha_{2,L-1}^{(L-1)}, 100 \cdots 0),$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$(\alpha_{q1}^{(L-1)}, \cdots \alpha_{q,L-1}^{(L-1)}, q-1, 00 \cdots 0).$$

At *Step* L, therefore, there remains a single comparison among q paths, whose survivor will be accepted as the correct path. While this decoding algorithm is clearly suboptimal, the optimal being a comparison of the likelihood functions of all q^L paths at the end of the tree based on $(L+K-1)v$ received channels symbols, we shall show in the next section that the algorithm is asymptotically optimum for $R > R_0 = E_0(1)$ for all but pathological channels.

V. Random Coding Upper Bound

If we now assume that the matrix G is randomly selected with a uniform distribution from the ensemble of q^{vK} matrices of elements in $GF(q)$ and the sequence $\mathbf{c}$ is also randomly selected from among all possible $(L+K-1)v$-dimensional vectors with components in the same field, the channel symbols along a given path regarded as random variables have the following properties[8] in addition to A):

B) The probability distribution of the jth channel symbol for any path is the same for all j, and for all paths

$$P(x_j = \xi_i) = P_i \qquad (i = 1, 2, \cdots r).$$

C) Successive channel symbols along a given path are statistically independent

$$P(x_1 = \xi_{i_1}, x_2 = \xi_{i_2}, \cdots x_{(L+K-1)v} = \xi_{i_{(L+K-1)v}}) = \prod_{j=1}^{(L+K-1)v} P(x_j = \xi_{i_j}).$$

We shall need one more property before we can proceed, which requires a modification of the encoder:

D) Symbols along arbitrary subsequences of any two totally distinct paths are independent.

Reiffen[8] proved property D) for the present encoder but only within the first K-branch constraint length. To ensure that D) is satisfied over the entire L-branch tree, we must modify the encoder. One obvious way is to randomly select a new Kxv generator matrix G after each new data symbol a_j is shifted into the register. However, Massey[9] has recently shown that it is possible to ensure D) by introducing only $2v$ new components into the first two rows of the generator matrix for each new data symbol, and simply shifting all the rows of the previous generator matrix two places downward and discarding the last two rows.

We now proceed to obtain an upper bound on the error probability for the class of convolutional codes which possess the above properties, by analyzing the performance of the decoding algorithm of the previous

section. We recall that the correct path is eliminated if it fails to have the largest likelihood function in any one of the L comparisons among q alternatives in which it is involved.

In particular, let us consider the situation at the $(j+1)$th step. Without loss of generality, we may assume that the correct path corresponds to the all zeros data sequence. Although the comparison at this step is with only $q-1$ other paths, there is a multitude of potential adversaries. Thus, with the first $j+K$ branches of the correct path denoted by the vector $\mathbf{0} = (00\cdots 0)$, consider all the paths of the form $\alpha_{21}^{(j)}, \alpha_{22}^{(j)}, \cdots \alpha_{2j}^{(j)}100\cdots 0$. There is only one such path which diverged from the correct path K branches back: namely, the one for which $\alpha_{21}^{(j)}\cdots\alpha_{2j}^{(j)} = 00\cdots 0$. But there are $q-1$ potential adversaries of this form which diverged from the correct path $K+1$ branches back: namely, those for which $\alpha_{21}^{(j)}\cdots\alpha_{2j-1}^{(j)} = 00\cdots 0$ and $\alpha_{2j}^{(j)}$ is any element of $GF(q)$ except 0. Similarly, there are $(q-1)q$ potential adversaries of this form which diverged from the correct path $K+2$ branches back: namely, those for which $\alpha_{21}^{(j)}\cdots\alpha_{2,j-2}^{(j)} = 00\cdots 0$, $\alpha_{2,j-1}^{(j)}$ is any element except 0, and $\alpha_{2j}^{(j)}$ is any element of $GF(q)$. Continuing in this way, we find that there are $(q-1)q^{l-1}$ potential adversaries of this form which diverged $K+l$ branches back. However, there are exactly as many potential adversaries for which $a_{j+1} = 2$, as these are adversaries for which $a_{j+1} = 1$, and similarly for $a_{j+1} = 3, 4, \cdots q-1$. Thus, the total number of potential adversaries which diverged from the correct path $K+l$ branches back ($l = 1, 2, \cdots$) is $(q-1)^2q^{l-1}$, while $q-1$ paths diverged K branches back.

Before we can proceed to bound the error probability, we must establish that of all the potential adversaries which diverged from the correct path $K+l$ branches back only those that are totally distinct from it can actually be adversaries in the comparison of likelihood functions. We recall from property A) that two paths which diverge at a given branch will converge again after K branches if all of the next K data symbols are identical. Furthermore, any pair of paths having data symbols which are never identical for K consecutive branches remain totally distinct from the initial divergent branch. We now observe that by the nature of the decoding algorithm no two adversaries in any comparison can agree in K (or more) consecutive branch data symbols beyond their point of initial divergence, for at the outcome of each preceding set of comparisons there was one and only one surviving path with a particular sequence of K data symbols.

Thus, all the actual adversaries to the correct path at step $j+1$ are totally distinct from it and consequently the branch channel symbols are statistically independent [Property D]. Further, we have no more than $q-1$ possible adversaries to the correct path which diverged K branches (or N channel symbols) back and no more than $(q-1)^2q^{l-1}$ possible adversaries to the correct path which diverged $K+l$ branches (or $(K+l)v = N + (\ln q/R)l$ channel symbols) back, where $l = 1, 2, \cdots$.

Thus, the expected probability of an error in the comparison at the $(j+1)$th step is bounded by the union bound,

$$\overline{P(j+1)} < \sum_{l=0}^{j} \overline{\Pr(\text{error caused by a possible adversary which diverged } K+l \text{ branches back})}. \quad (19)$$

The zeroth term of this sum is bounded by the probability of error for a block code of $(q-1)$ words (the maximum number of possible adversaries) each of length N channel symbols, while the lth term ($l \geq 1$) is bounded by the error probability for a block code of $(q-1)^2q^{l-1}$ words each of length $N + (\ln q/R)l$ channel symbols. Since all symbols of each codeword are mutually independent and symbols of the correct codeword are independent of symbols of any other codeword, we may use the random coding upper bound on block codes[5],[1] for the lth term. Thus, if for the given transmission rate the convolutional encoder is mechanized, as described above, so that the input symbol distribution is that which achieves the maximum of (7), we have,

$$\begin{aligned}\overline{P(j+1)} &< (q-1)^{\rho}\exp[-NE_0(\rho)] + \sum_{l=1}^{j}[(q-1)^2q^{l-1}]^{\rho} \\ &\quad\cdot\exp\left[-\left(N+\frac{\ln q}{R}l\right)E_0(\rho)\right] \\ &< (q-1)\exp[-NE_0(\rho)]\sum_{l=0}^{\infty}q^{l[\rho-E_0(\rho)/R]} \\ &= \frac{q-1}{1-q^{-\epsilon/R}}\exp[-NE_0(\rho)] \quad (0<\rho\leq 1) \quad (20)\end{aligned}$$

where $\epsilon = E_0(\rho) - \rho R > 0$. This bound is independent of j. We again use a union bound to express the error probability in decoding the L branch tree in terms of (20) and thus obtain

$$\begin{aligned}\overline{P_E} &< \sum_{j=0}^{L-1}\overline{P(j+1)} \\ &< \frac{L(q-1)}{1-q^{-\epsilon/R}}\exp[-NE_0(\rho)] \quad (0<\rho\leq 1)\end{aligned} \quad (21)$$

where $\epsilon = E_0(\rho) - \rho R > 0$ and since at least one code in the ensemble must have $P_E < \overline{P_E}$, and $E_0(\rho)$ is a monotonically increasing function of ρ, we have

Theorem 2

The probability of error in decoding an L-branch q-ary tree code transmitted over a memoryless channel is bounded by

$$P_E < \frac{L(q-1)}{1-q^{-\epsilon/R}}\exp[-NE(R)]$$

[1] Note that Gallager's proof of the upper bound for block codes[5] requires only that the correct word symbols be independent of the symbols of any incorrect word, and not that incorrect words be mutually independent.

where[2]

$$E(R) = \begin{cases} R_0, & 0 \le R = R_0 - \epsilon < R_0 \quad (22a) \\ E_0(\rho), & R_0 - \epsilon \le R = \dfrac{E_0(\rho) - \epsilon}{\rho} < C \\ & (0 < \rho \le 1) \quad (22b) \end{cases}$$

and

$$R_0 = E_0(1) = \max_{p(x)} \{-\ln \sum_Y [\sum_X p(x)\sqrt{p(y \mid x)}]^2\}.$$

Since the bound was shown for the specific probabilistic decoding algorithm described above, and $\epsilon > 0$ can be made arbitrarily small for N arbitrarily large, we have comparing (16) and (22), whenever $E_0(\rho)$ is concave,

$$\lim_{N\to\infty} \frac{\ln (1/P_E)}{N} = E(R) = E_L(R) \quad \text{for} \quad R_0 \le R < C \quad (23)$$

and consequently

Corollary 1

For all but pathological channels the specific probabilistic decoding algorithm described in Section IV is asymptotically (exponentially) optimum for $R \ge R_0$.

Yudkin[2] has obtained an upper bound with the exponent of (22) for the undetectable error probability of the Fano sequential decoding algorithm.[3] Thus the Fano algorithm is also asymptotically optimum in this sense for $R \ge R_0$. However, the average number of computations per branch is unbounded for $R > R_0$ in the latter, while for the nonsequential algorithm considered here the number of computations per branch is proportional to q^K independent of rate. Also, as we shall show below, the number of computations required with this algorithm for a convolutional code of constraint length N is essentially the same as the number required by a maximum likelihood decoder for a block code of block length N, all the other parameters being the same.

The random coding upper bound exponent (with $\epsilon = 0$) is greater than the random coding exponent for block codes for all rates ($0 < R < C$), as is seen by comparing (22) with the exponent for block codes[5] of length N:

$$E(R) = \begin{cases} R_0 - R, & 0 \le R \le E_0'(1) \quad (24a) \\ E_0(\rho) - \rho E_0'(\rho), & E_0'(1) \le R = E_0'(\rho) < C \\ & (0 < \rho \le 1). \quad (24b) \end{cases}$$

From property b) of $E_0(\rho)$, we have $E_0'(\rho) > 0$. Also, from (24b) we have $E_0(\rho)/\rho \ge E_0'(\rho)$, and the conclusion follows.

The same is true also for the lower bound. For $R > E_0'(\tilde{\rho})$, the best known lower bound for block codes[5]-[7] coincides with the sphere packing bound, which is the same as (24b) for nonpathological channels but with ρ extended to $\tilde{\rho} \ge 1$. Thus for this range the lower bound on convolutional codes (16) exceeds this for the reasons just stated. For $R < E_0'(\tilde{\rho})$, the best known bound for block codes[7] is $E_L(R) = E_x - \tilde{\rho}R$ ($\rho \ge 1$), while from (18) for convolutional codes we have $E_L(R) = E_x$ for $0 < R \le E_0(\tilde{\rho})/\tilde{\rho} > E_0'(\tilde{\rho})$ which therefore exceeds the lower bound for block codes in this region also. For pathological channels the same argument applies to $\hat{E}_0(\rho)$.

[2] If $E_0''(\rho) > 0$ for some ρ on the unit interval, (22b) may specify more than one value of $E(R)$ for a given R. In this case we should choose the greater, with the result that $E(R)$ is a discontinuous function.

VI. Limiting Cases and Comparisons with Block Codes

Of particular interest is the behavior of the exponent in the neighborhood of capacity. We have from the properties a), b), and equation (7)

$$\hat{E}_0(0) = 0, \qquad \hat{E}_0'(0) = C, \qquad E_0''(0) \le 0.$$

We must solve the parametric equations

$$E_L(R) = \hat{E}_0(\rho) \quad (25a)$$

$$R = \frac{\hat{E}_0(\rho)}{\rho} \qquad (0 \le \rho \le 1) \quad (25b)$$

for R in the neighborhood of C, which corresponds to ρ in the neighborhood of 0. Thus, excluding for this purpose the case in which $E_0''(0) = 0$, and expanding $\hat{E}_0(\rho)$ in a Taylor series about $\rho = 0$ neglecting terms higher than quadratic, we obtain

$$\hat{E}_0(\rho) \approx \rho C + \frac{\rho^2}{2} E_0''(0) \approx E_0(\rho). \quad (26)$$

Then from (25b) and (26) we have

$$\rho = \frac{2(C - R)}{-E_0''(0)}.$$

Substituting in (26) and neglecting terms higher than linear in $C - R$ we obtain (setting $\epsilon \approx 0$ in the upper bound)

$$E(R) \approx E_L(R) = \hat{E}_0(\rho) \approx \frac{2C}{-E_0''(0)} (C - R).$$

In contrast, for block codes the exponent for rates in the neighborhood of $C(\rho = 0)$, as obtained by repeating the above argument in connection with (24b), is

$$E(R) = E_L(R) \approx \frac{1}{-2E_0''(0)} (C - R)^2.$$

Another interesting limiting case is that of "very noisy" channels which includes the time-discrete white Gaussian channel. A memoryless channel is said to be very noisy if $p(y \mid x) = p(y)(1 + \epsilon_{xy})$ where $|\epsilon_{xy}| \ll 1$ for all x and y in the channel input and output spaces X and Y. For this class of channels it has been shown[5] that when the input distribution is optimized so that $I(X; Y) = C$, then

$$\hat{E}_0(\rho) = E_0(\rho) \approx \frac{\rho C}{1 + \rho} \quad (27)$$

where

$$C \approx \max_{p(x)} \sum_X \sum_Y p(x)p(y) \frac{\epsilon_{xy}^2}{2}.$$

Also

$$R_0 = E_0(1) \approx \frac{C}{2} \approx E_x$$

and from (17b), it follows that $\tilde{\rho} = 1$. Thus, with $\epsilon = 0$ we find from (18), (22), and (27)

$$E(R) \approx E_L(R) \approx C/2, \qquad 0 \leq R \leq C/2. \tag{28a}$$

For rates above $C/2$ we have from (16), (22), and (27)

$$R = \frac{E_0(\rho)}{\rho} \approx \frac{C}{1+\rho}.$$

Solving for ρ in terms of R, and substituting in (27), we obtain from (16) and (22):

$$E(R) \approx E_L(R) \approx C - R, \qquad \frac{C}{2} \leq R < C. \tag{28b}$$

From (28a) and (28b) we note that for very noisy channels the upper and lower bounds are exponentially equal for all rates, that they remain at the zero rate level of $C/2$ up to $R = C/2$ and then decrease linearly for rates up to C. This is to be compared with the corresponding result for block codes:[5]

$$E(R) \approx E_L(R) \approx \begin{cases} \frac{C}{2} - R, & 0 \leq R \leq C/4 \\ (\sqrt{C} - \sqrt{R})^2, & C/4 \leq R < C. \end{cases} \tag{29}$$

The two exponents for very noisy channels (28) and (29) are plotted in Fig. 5. The relative improvement increases with rate. For $R = R_0 = C/2$, the exponent for convolutional codes is almost six times that for block codes.

While the upper and lower bound exponents are identical in the limiting case, we see from the example of the error-bound exponents for three binary symmetric channels (with $p = 0.01$, $p = 0.1$, and $p = 0.4$), shown normalized by C in Fig. 6, that as the channel becomes less noisy the upper and lower bounds diverge for $R < R_0$. In fact, if for all ρ, $E_0''(\rho) \equiv 0$, then $E_0(\rho) = \rho C$, so that $R_0 = C$. Thus, the upper bound exponent equals R_0 for all $R < C$.

There remains to show that this significant improvement over the performance of block codes is achievable without additional decoding complexity. But we observe that in decoding L branches or $L \ln q$ nats the decoding algorithm considered makes slightly less than Lq^K branch likelihood function computations or $Lvq^K = (L/K)Nq^K$ symbol likelihood function computations. Now the equivalent block code transmits $L \ln q$ nats in blocks of $K \ln q$ nats at a rate $R = \ln q/v = K \ln q/N$ nats/symbol, which corresponds to transmitting one of q^K words of length N symbols. Thus, the decoder must perform Nq^K symbol likelihood function computations per block and

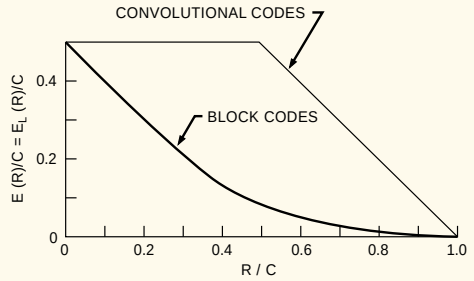


Fig. 5. *E (R)* for very noisy channels with convolutional and block codes.

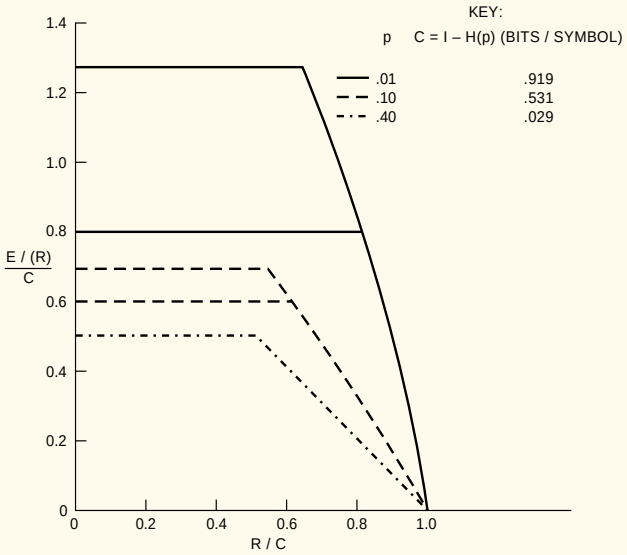


Fig. 6. *E (R)* and $E_L(R)$ for the binary symmetric channels with evolutional codes ($p = 0.01$, $p = 0.1$, $p = 0.4$).

repeat this L/K times. Consequently, the number of computations is essentially the same for the convolutional code decoding algorithm described as is required for maximum likelihood decoding of the equivalent block code.

We should note, however, that since $K - 1$ zeros are inserted between trees of L branches, the actual rate for convolution codes is reduced by a factor of $L/(L + K - 1)$ from that of block codes, a minor loss since, because of the greatly increased exponent, we can afford to increase L (which affects P_E only linearly) enough to make this factor insignificant.

VII. A Semi-Sequential Modification of the Decoding Algorithm

We observe from (22) with the substitution $N = Kv = K \ln q/R$, that

$$P_E < \frac{L(q-1)}{1 - q^{-\epsilon/R}} (q^K)^{-R_0/R} \quad \text{for } 0 \leq R = R_0 - \epsilon < R_0 \tag{30}$$

for the specific decoding algorithm considered. However, as we have just noted, the number of likelihood function computations per decoded branch is slightly less that q^K,

which means that the error probability decreases more than linearly with computational complexity for rates in this region.

Now let us consider an iterated version of the previous algorithm. At first we shall employ the aid of a magic genie. It is clear that the nonsequential decoding algorithm can be modified to make decisions based on k branches where $k < K$, the constraint length, and that the resulting error probability is the same as (30) with K replaced by k. Thus suppose the decoder attempts to decode the L-branch tree using $k = 1$ and at the end of the tree the genie either tells him he is correct or requires him to start over with $k = 2$ and that he proceeds in this way each time increasing k by 1 until he is either told he is correct or he reaches the constraint length K. Then, since at each iteration the number of computations is increased by a factor q, the number of computations per branch performed by the end of the kth iteration is $q + q^2 + \cdots + q^k = [q(q^k - 1)/(q - 1)] < 2q^k$. Thus, denoting the total number of computations per branch by γ, we have using (30),

$$\text{Prob } (\gamma > 2q^k) < \frac{L(q-1)}{1 - q^{-\epsilon/R}} (q^k)^{-R_0/R},$$

$$0 \leq R = R_0 - \epsilon < R_0$$

or

$$\text{Prob } (\gamma > \Gamma) < \frac{L(q-1)}{1 - q^{-\epsilon/R}} \left(\frac{\Gamma}{2}\right)^{-R_0/R},$$

$$0 \leq R = R_0 - \epsilon < R_0 \qquad (31)$$

which is known as a Pareto distribution. Also, we have for the expected number of computations per branch

$$\bar{\gamma} < \sum_{k=1}^{K} q^k P_E(k-1) < \frac{L(q-1)}{1 - q^{-\epsilon/R}} \sum_{k=1}^{\infty} q^{-[(\epsilon k - R_0)/R]}$$

$$= \frac{L(q-1)q}{(1 - q^{-\epsilon/R})^2}, \qquad 0 \leq R = R_0 - \epsilon < R_0. \qquad (32)$$

Thus, the expected number of computations per branch increases no more rapidly than the tree length for $R < R_0$, a feature of sequential decoding. Actually the Fano algorithm has been shown[10] to have a Pareto distribution on the number of computations with a higher exponent than R_0/R for $R < R_0$ and an expected number of computations which is independent of the tree or constraint length. However, with the Wozencraft algorithm[4] $\bar{\gamma}$ increases linearly with constraint length. The major drawback of this scheme, besides the genie which we shall dispose of presently, is that the number of storage registers required at the kth iteration is q^k and consequently the required storage capacity also has a Pareto distribution.

To avoid employing the genie, the decoder must have some other way to decide whether or not the kth iteration produces the correct path. One way to achieve this is to compare the likelihood function for the last N symbols of the decoded path with a threshold. If it exceeds this threshold the total path is accepted as correct; otherwise the algorithm is repeated with k increased by 1. Since the last N symbols occur after the tree has stopped branching, these can be affected by the last K branches only since no more than K data symbols are in the coder shift register when these channel symbols are being generated. Thus, there are only q^K possible combinations of channel symbols for the final branches which are of length N channel symbols. The upper bound on the probability of error for a threshold decision involving q^K code words of block length N selected independently is[11]

$$P_T < 2 \exp [-NE_T(R)]$$

where

$$E_T(R) = \max_{p(x)} \{\max_{0 \leq \rho \leq 1} \cdot [-\ln \sum_X \sum_Y p(x)p(y \mid x)^{1-\rho} p(y)^{\rho} - \rho R]\} > 0,$$

$$0 \leq R < C$$

and

$$R = \frac{K \ln q}{N} = \frac{\ln q}{v} \quad \text{as before.}$$

By choosing N or K large enough, P_T can be made sufficiently small, although clearly it can not be as small as P_E of (22), which results from use of the nonsequential algorithm.

Although this algorithm is rendered impractical by the excessive storage requirements, it contributes to a general understanding of convolutional codes and sequential decoding through its simplicity of mechanization and analysis.

Acknowledgment

The author gratefully acknowledges the helpful suggestions and patience of Dr. L. Kleinrock during numerous discussions.

References

[1] P. Elias, "Coding for noisy channels," *IRE Conv. Rec.*, pt. IV, pp. 37–46, 1955.

[2] H. L. Yudkin, "Channel state testing in information decoding," Ph.D. dissertation, Dept. of Elec. Engrg., M.I.T., Cambridge, Mass., September 1964.

[3] R. M. Fano, "A heuristic discussion of probabilistic decoding," *IEEE Trans. on Information Theory*, vol. IT-9, pp. 64–76, April 1963.

[4] J. M. Wozencraft and B. Reiffen, *Sequential Decoding*. Cambridge, Mass.: M.I.T. Press, and New York: Wiley, 1961.

[5] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. on Information Theory*, vol. IT-11, pp. 3–18, January 1965.

[6] R. M. Fano, *Transmission of Information*. Cambridge, Mass.: M.I.T. Press, and New York: Wiley, 1961.

[7] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels," *Information and Control* (to be published).

[8] B. Reiffen, "Sequential encoding and decoding for the discrete memoryless channel," M.I.T. Lincoln Laboratory, Lexington, Mass., Rept. 25, *G*-0018, August 1960.

[9] J. L. Massey, private communication.

[10] J. E. Savage, "Sequential decoding—The computation problem," *Bell Sys. Tech. J.*, vol. 45, pp. 149–175, January 1966.

[11] C. E. Shannon, unpublished notes.

IEEE TRANSACTIONS ON

COMMUNICATION TECHNOLOGY

OCTOBER 1971 VOLUME COM-19 NUMBER 5

Convolutional Codes and Their Performance in Communication Systems

ANDREW J. VITERBI, SENIOR MEMBER, IEEE

Abstract—**This tutorial paper begins with an elementary presentation of the fundamental properties and structure of convolutional codes and proceeds with the development of the maximum likelihood decoder. The powerful tool of generating function analysis is demonstrated to yield for arbitrary codes both the distance properties and upper bounds on the bit error probability for communication over any memoryless channel. Previous results on code ensemble average error probabilities are also derived and extended by these techniques. Finally, practical considerations concerning finite decoding memory, metric representation, and synchronization are discussed.**

I. Introduction

ALTHOUGH convolutional codes, first introduced by Elias [1], have been applied over the past decade to increase the efficiency of numerous communication systems, where they invariably outperform block codes of the same order of complexity, there remains to date a lack of acceptance of convolutional coding and decoding techniques on the part of many communication technologists. In most cases, this is due to an incomplete understanding of convolutional codes, whose cause can be traced primarily to the sizable literature in this field, composed largely of papers which emphasize details of the decoding algorithms rather than the more fundamental unifying concepts, and which, until recently, have been divided into two nearly disjoint subsets. This malady is shared by the block-coding literature, wherein the algebraic decoders and probabilistic decoders have been at odds for a considerably longer period.

The convolutional code dichotomy owes its origins to the development of sequential (probabilistic) decoding by Wozencraft [2] and of threshold (feedback, algebraic) decoding by Massey [3]. Until recently the two disciplines flourished almost independently, each with its own literature, applications, and enthusiasts. The Fano sequential decoding algorithm [4] was soon found to

Paper approved by the Communication Theory Committee of the IEEE Communication Technology Group for publication without oral presentation. Manuscript received January 7, 1971; revised June 11, 1971.

The author is with the School of Engineering and Applied Science, University of California, Los Angeles, Calif. 90024, and the Linkabit Corporation, San Diego, Calif.

greatly outperform earlier versions of sequential decoders both in theory and practice. Meanwhile the feedback decoding advocates were encouraged by the burst-error correcting capabilities of the codes which render them quite useful for channels with memory.

To add to the confusion, yet a third decoding technique emerged with the Viterbi decoding algorithm [9], which was soon thereafter shown to yield maximum likelihood decisions (Forney [12], Omura [17]). Although this approach is probabilistic and emerged primarily from the sequential-decoding oriented discipline, it leads naturally to a more fundamental approach to convolutional code representation and performance analysis. Furthermore, by emphasizing the decoding-invariant properties of convolutional codes, one arrives directly to the maximum likelihood decoding algorithm and from it to the alternate approaches which lead to sequential decoding on the one hand and feedback decoding on the other. This decoding algorithm has recently found numerous applications in communication systems, two of which are covered in this issue (Heller and Jacobs [24], Cohen *et al.* [25]). It is particularly desirable for efficient communication at very high data rates, where very low error rates are not required, or where large decoding delays are intolerable.

Foremost among the recent works which seek to unify these various branches of convolutional coding theory is that of Forney [12], [21], [22], *et seq.*, which includes a three-part contribution devoted, respectively, to algebraic structure, maximum likelihood decoding, and sequential decoding. This paper, which began as an attempt to present the author's original paper [9] to a broader audience,[1] is another such effort at consolidating this discipline.

It begins with an elementary presentation of the fundamental properties and structure of convolutional codes and proceeds to a natural development of the maximum likelihood decoder. The relative distances among codewords are then determined by means of the generating function (or transfer function) of the code state diagram. This in turn leads to the evaluation of coded communication system performance on any memoryless channel. Performance is first evaluated for the specific cases of the binary symmetric channel (BSC) and the additive white Gaussian noise (AWGN) channel with biphase (or quadriphase) modulation, and finally generalized to other memoryless channels. New results are obtained for the evaluation of specific codes (by the generating function technique), rather than the ensemble average of a class of codes, as had been done previously, and for bit error probability, as distinguished from event error probability.

The previous ensemble average results are then extended to bit error probability bounds for the class of time-varying convolutional codes by means of a generalized generating function approach; explicit results are obtained for the limiting case of a very noisy channel and compared with the corresponding results for block codes. Finally, practical considerations concerning finite memory, metric representation, and synchronization are discussed. Further and more explicit details on these problems and detailed results of performance analysis and simulation are given in the paper by Heller and Jacobs [24].

While sequential decoding is not treated explicitly in this paper, the fundamentals and techniques presented here lead naturally to an elegant tutorial presentation of this subject, particularly if, following Jelinek [18], one begins with the recently proposed stack sequential decoding algorithm proposed independently by Jelinek and Zigangirov [7], which is far simpler to describe and understand then the original sequential algorithms. Such a development, which proceeds from maximum likelihood decoding to sequential decoding, exploiting the similarities in performance and analysis has been undertaken by Forney [22]. Similarly, the potentials and limitations of feedback decoders can be better understood with the background of the fundamental decoding-invariant convolutional code properties previously mentioned, as demonstrated, for example, by the recent work of Morrissey [15].

II. Code Representation

A convolutional encoder is a linear finite-state machine consisting of a K-stage shift register and n linear algebraic function generators. The input data, which is usually, though not necessarily, binary, is shifted along the register b bits at a time. An example with $K = 3$, $n = 2, b = 1$ is shown in Fig. 1.

The binary input data and output code sequences are indicated on Fig. 1. The first three input bits, 0, 1, and 1, generate the code outputs 00, 11, and 01, respectively. We shall pursue this example to develop various representations of convolutional codes and their properties. The techniques thus developed will then be shown to generalize directly to any convolutional code.

It is traditional and instructive to exhibit a convolutional code by means of a tree diagram as shown in Fig. 2.

If the first input bit is a zero, the code symbols are those shown on the first upper branch, while if it is a one, the output code symbols are those shown on the first lower branch. Similarly, if the second input bit is a zero, we trace the tree diagram to the next upper branch, while if it is a one, we trace the diagram downward. In this manner all 32 possible outputs for the first five inputs may be traced.

From the diagram it also becomes clear that after the first three branches the structure becomes repetitive. In fact, we readily recognize that beyond the third branch the code symbols on branches emanating from the two nodes labeled a are identical, and similarly for all the

[1] This material first appeared in unpublished form as the notes for the Linkabit Corp., "Seminar on convolutional codes," Jan. 1970.

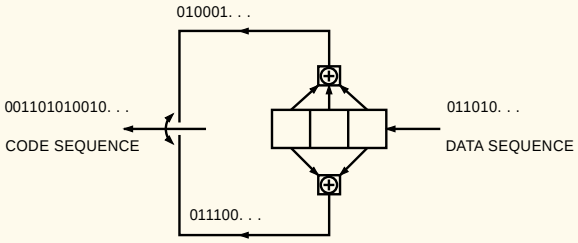


Fig. 1. Convolutional coder for $K = 3$, $n = 2$, $b = 1$.

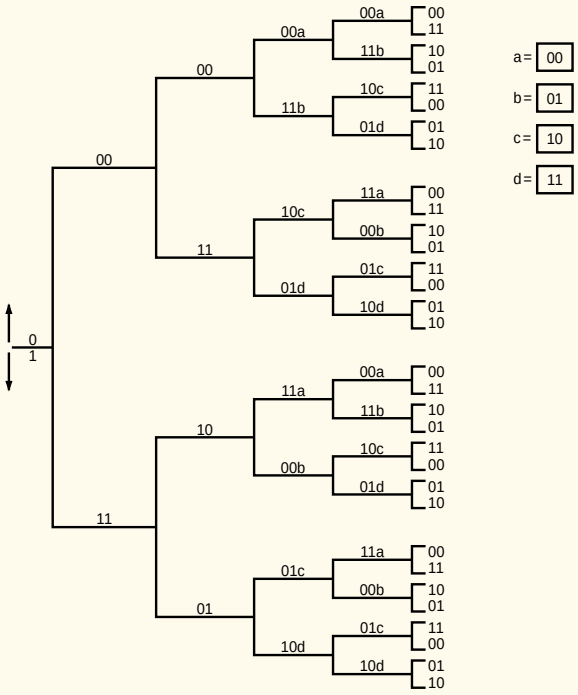


Fig. 2. Tree-code representation for coder of Fig. 1.

identically labeled pairs of nodes. The reason for this is obvious from examination of the encoder. As the fourth input bit enters the coder at the right, the first data bit falls off on the left end and no longer influences the output code symbols. Consequently, the data sequences $100xy\cdots$ and $000xy\cdots$ generate the same code symbols after the third branch and, as is shown in the tree diagram, both nodes labeled a can be joined together.

This leads to redrawing the tree diagram as shown in Fig. 3. This has been called a trellis diagram [12], since a trellis is a tree-like structure with remerging branches. We adopt the convention here that code branches produced by a "zero" input bit are shown as solid lines and code branches produced by a "one" input bit are shown dashed.

The completely repetitive structure of the trellis diagram suggests a further reduction in the representation of the code to the state diagram of Fig. 4. The "states" of the state diagram are labeled according to the nodes of the trellis diagram. However, since the states correspond merely to the last two input bits to the coder we may use these bits to denote the nodes or states of this diagram.

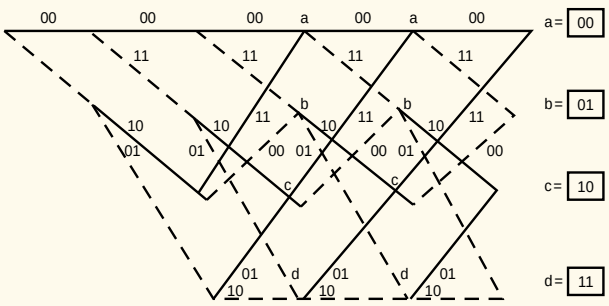


Fig. 3. Trellis-code representation for coder of Fig. 1.

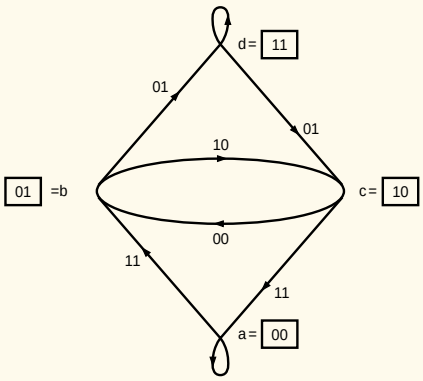


Fig. 4. State-diagram representation for coder of Fig. 1.

We observe finally that the state diagram can be drawn directly by observing the finite-state machine properties of the encoder and particularly the fact that a four-state directed graph can be used to represent uniquely the input–output relation of the eight-state machine. For the nodes represent the previous two bits while the present bit is indicated by the transition branch; for example, if the encoder (machine) contains 011, this is represented in the diagram by the transition from state $b = 01$ to state $d = 11$ and the corresponding branch indicates the code symbol outputs 01.

III. Minimum Distance Decoder for Binary Symmetric Channel

On a BSC, errors which transform a channel code symbol 0 to 1 or 1 to 0 are assumed to occur independently from symbol to symbol with probability p. If all input (message) sequences are equally likely, the decoder which minimizes the overall error probability for any code, block or convolutional, is one which examines the error-corrupted received sequence $y_1y_2\cdots y_j\cdots$ and chooses the data sequence corresponding to the transmitted code sequence $x_1x_2\cdots x_j\cdots$, which is closest to the received sequence in the sense of Hamming distance; that is, the transmitted sequence which differs from the received sequence in the minimum number of symbols.

Referring first to the tree diagram, this implies that we should choose that path in the tree whose code sequence differs in the minimum number of symbols from the received sequence. However, recognizing that the transmitted code branches remerge continually, we may equally limit our choice to the possible paths in the trellis diagram of Fig. 3. Examination of this diagram indicates that it is unnecessary to consider the entire received sequence (which conceivably could be thousands or millions of symbols in length) at one time in deciding upon the most likely (minimum distance) transmitted sequence. In particular, immediately after the third branch we may determine which of the two paths leading to node or state a is more likely to have been sent. For example, if 010001 is received, it is clear that this is at distance 2 from 000000 while it is at distance 3 from 111011 and consequently we may exclude the lower path into node a. For, no matter what the subsequent received symbols will be, they will effect the distances only over subsequent branches after these two paths have remerged and consequently in exactly the same way. The same can be said for pairs of paths merging at the other three nodes after the third branch. We shall refer to the minimum distance path of the two paths merging at a given node as the "survivor." Thus it is necessary only to remember which was the minimum distance path from the received sequence (or survivor) at each node, as well as the value of that minimum distance. This is necessary because at the next node level we must compare the two branches merging at each node level, which were survivors at the previous level for different nodes; e.g., the comparison at node a after the fourth branch is among the survivors of comparisons at nodes a and c after the third branch. For example, if the received sequence over the first four branches is 01000111, the survivor at the third node level for node a is 000000 with distance 2 and at node c it is 110101, also with distance 2. In going from the third node level to the fourth the received sequence agrees precisely with the survivor from c but has distance 2 from the survivor from a. Hence the survivor at node a of the fourth level is the data sequence 1100 which produced the code sequence 11010111 which is at (minimum) distance 2 from the received sequence.

In this way we may proceed through the received sequence and at each step for each state preserve one surviving path and its distance from the received sequence, which is more generally called *metric*. The only difficulty which may arise is the possibility that in a given comparison between merging paths, the distances or metrics are identical. Then we may simply flip a coin as is done for block codewords at equal distances from the received sequence. For even if we preserved both of the equally valid contenders, further received symbols would affect both metrics in exactly the same way and thus not further influence our choice.

This decoding algorithm was first proposed by Viterbi [9] in the more general context of arbitrary memoryless channels. Another description of the algorithm can be obtained from the state-diagram representation of Fig. 4. Suppose we sought that path around the directed state diagram, arriving at node a after the kth transition, whose code symbols are at a minimum distance from the received sequence. But clearly this minimum distance path to node a at time k can be only one of two candidates: the miminum distance path to node a at time $k-1$ and the minimum distance path to node c at time $k-1$. The comparison is performed by adding the new distance accumulated in the kth transition by each of these paths to their minimum distances (metrics) at time $k-1$.

It appears thus that the state diagram also represents a system diagram for this decoder. With each node or state we associate a storage register which remembers the minimum distance path into the state after each transition as well as a metric register which remembers its (minimum) distance from the received sequence. Furthermore, comparisons are made at each step between the two paths which lead into each node. Thus four comparators must also be provided.

There remains only the question of truncating the algorithm and ultimately deciding on one path rather than four. This is easily done by forcing the last two input bits to the coder to be 00. Then the final state of the code must be $a = 00$ and consequently the ultimate survivor is the survivor at node a, after the insertion into the coder of the two dummy zeros and transmission of the corresponding four code symbols. In terms of the trellis diagram this means that the number of states is reduced from four to two by the insertion of the first zero and to a single state by the insertion of the second. The diagram is thus truncated in the same way as it was begun.

We shall proceed to generalize these code representations and optimal decoding algorithm to general convolutional codes and arbitrary memoryless channels, including the Gaussian channel, in Sections V and VI. However, first we shall exploit the state diagram further to determine the relative distance properties of binary convolutional codes.

IV. Distance Properties of Convolutional Codes

We continue to pursue the example of Fig. 1 for the sake of clarity; in the next section we shall easily generalize results. It is well known that convolutional codes are group codes. Thus there is no loss in generality in computing the distance from the all zeros codeword to all the other codewords, for this set of distances is the same as the set of distances from any specific codeword to all the others.

For this purpose we may again use either the trellis diagram or the state diagram. We first of all redraw the trellis diagram in Fig. 5 labeling the branches according to their distances from the all zeros path. Now consider all the paths that merge with the all zeros for the first time at some arbitrary node j.

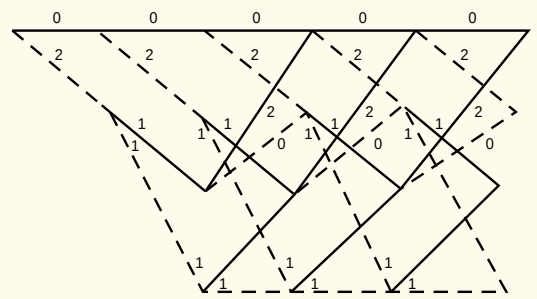


Fig. 5. Trellis diagram labeled with distances from all zeros path.

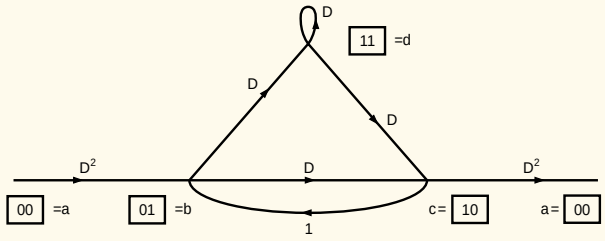


Fig. 6. State diagram labeled according to distance from all zeros path

It is seen from the diagram that of these paths there will be just one path at distance 5 from the all zeros path and this diverged from it three branches back. Similarly there are two at distance 6 from it, one which diverged 4 branches back and the other which diverged 5 branches back, and so forth. We note also that the input bits for distance 5 path are 00 ··· 0100 and thus differ in only one input bit from the all zeros, while the distance 6 paths are 00 ··· 01100 and 00 ··· 010100 and thus each differs in 2 input bits from the all zeros path. The minimum distance, sometimes called the minimum "free" distance, among all paths is thus seen to be 5. This implies that any pair of channel errors can be corrected, for two errors will cause the received sequence to be at distance 2 from the transmitted (correct) sequence but it will be at least at distance 3 from any other possible code sequence. It appears that with enough patience the distance of all paths from the all zeros (or any arbitrary) path can be so determined from the trellis diagram.

However, by examining instead the state diagram we can readily obtain a closed form expression whose expansion yields directly and effortlessly all the distance information. We begin by labeling the branches of the state diagram of Fig. 4 either D^2, D, or $D^0 = 1$, where the exponent corresponds to the distance of the particular branch from the corresponding branch of the all zeros path. Also we split open the node $a = 00$, since circulation around this self-loop simply corresponds to branches of the all zeros path whose distance from itself is obviously zero. The result is Fig. 6. Now as is clear from examination of the trellis diagram, every path which arrives at state $a = 00$ at node level j, must have at some previous node level (possibly the first) originated at this same state $a = 00$. All such paths can be traced on the modified state diagram. Adding branch exponents we see that path $a\ b\ c\ a$ is at distance 5 from the correct path, paths $a\ b\ d\ c\ a$ and $a\ b\ c\ b\ c\ a$ are both at distance 6, and so forth, for the *generating functions* of the output sequence weights of these paths are D^5 and D^6, respectively

Now we may evaluate the generating function of all paths merging with the all zeros at the jth node level simply by evaluating the generating function of all the weights of the output sequences of the finite-state machine.[2] The result in this case is

$$T(D) = \frac{D^5}{1 - 2D}$$
$$= D^5 + 2D^6 + 4D^7 + \cdots + 2^k D^{k+5} + \cdots . \quad (1)$$

This verifies our previous observation and in fact shows that among the paths which merge with the all zeros at a given node there are 2^k paths at distance $k + 5$ from the all zeros.

Of course, (1) holds for an infinitely long code sequence; if we are dealing with the jth node level, we must truncate the series at some point. This is most easily done by considering the additional information indicated in the modified state diagram of Fig. 7.

The L terms will be used to determine the length of a given path; since each branch has an L, the exponent of the L factor will be augmented by one every time a branch is passed through. The N term is included only if that branch transition was caused by an input data "one," corresponding to a dotted branch in the trellis diagram. The generating function of this augmented state diagram is then

$$T(D, L, N)$$
$$= \frac{D^5L^3N}{1 - DL(1 + L)N}$$
$$= D^5L^3N + D^6L^4(1 + L)N^2 + D^7L^5(1 + L)^2N^3$$
$$+ \cdots + D^{5+k}L^{3+k}(1 + L)^kN^{1+k} + \cdots . \quad (2)$$

Thus we have verified that of the two distance 6 paths one is of length 4 and the other is of length 5 and both differ in 2 input bits from the all zeros.[3] Also, of the distance 7 paths, one is of length 5, two are of length 6, and one is of length 7; all four paths correspond to input sequences with three ones. If we are interested in the jth node level, clearly we should truncate the series such that no terms of power greater than L^j are included.

We have thus fully determined the properties of all paths in the convolutional code. This will be useful later in evaluating error probability performance of codes used over arbitrary memoryless channels.

[2] Alternatively, this can be regarded as the transfer function of the diagram regarded as a signal flow graph.

[3] Thus if the all zeros was the correct path and the noise causes us to choose one of the incorrect paths, two bit errors will be made.

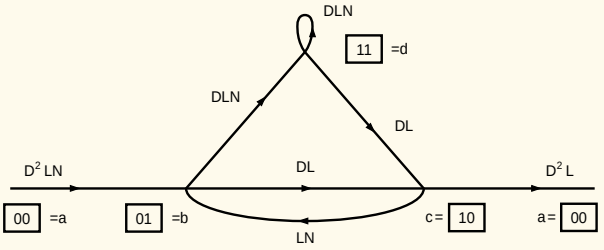


Fig. 7. State diagram labeled according to distance, length and number of input ones

V. Generalization to Arbitrary Convolutional Codes

The generalization of these techniques to arbitrary binary-tree ($b = 1$) convolutional codes is immediate. That is, a coder with a K-stage shift register and n mod-2 adders will produce a trellis or state diagram with 2^{K-1} nodes or states and each branch will contain n code symbols. The rate of this code is then

$$R = \frac{1}{n} \text{ bits/code symbol.}$$

The example pursued in the previous sections had rate $R = 1/2$. The primary characteristic of the binary-tree codes is that only two branches exit from and enter each node.

If rates other than $1/n$ are desired we must make $b > 1$, where b is the number of bits shifted into the register at one time. An example for $K = 2$, $b = 2$, $n = 3$, and consequently rate $R = 2/3$ is shown in Fig. 8 and its state diagram is shown in Fig. 9. It differs from the binary-tree codes only in that each node is connected to four other nodes, and for general b it will be connected to 2^b nodes. Still all the preceding techniques including the trellis and state-diagram generating function analysis are still applicable. It must be noted, however, that the minimum distance decoder must make comparisons among all the paths entering each node at each level of the trellis and select one survivor out of four (or out of 2^b in general).

VI. Generalization of Optimal Decoder to Arbitrary Memoryless Channels

Fig. 10 exhibits a communication system employing a convolutional code. The convolutional encoder is precisely the device studied in the preceding sections. The data sequence is generally binary ($a_i = 0$ or 1) and the code sequence is divided into subsequences where $\mathbf{x}_j$ represents the n code symbols generated just after the input bit a_i enters the coder: that is, the symbols of the jth branch. In terms of the example of Fig. 1, $a_3 = 1$ and $\mathbf{x}_3 = 01$. The channel output or received sequence is similarly denoted. $\mathbf{y}_i$ represents the n symbols received when the n code symbols of $\mathbf{x}_j$ were transmitted. This model includes the BSC wherein the $\mathbf{y}_j$ are binary n vectors each of whose symbols differs from the corresponding symbol of $\mathbf{x}_j$ with probability p and is identical to it with probability $1 - p$.

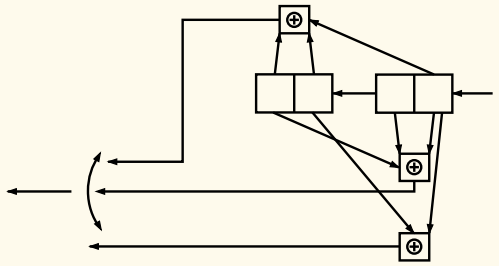


Fig. 8. Coder for $K = 2$, $b = 2$, $n = 3$, and $R = 2/3$.

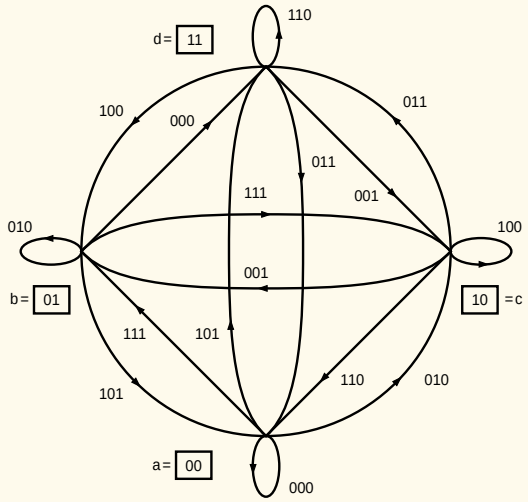


Fig. 9. State diagram for code of Fig. 8.

For completely general channels it is readily shown [6], [14] that if all input data sequences are equally likely, the decoder which minimizes the error probability is one which compares the conditional probabilities, also called likelihood functions, $P(\mathbf{y} \mid \mathbf{x}^{(m)})$, where $\mathbf{y}$ is the overall received sequence and $\mathbf{x}^{(m)}$ is one of the possible transmitted sequences, and decides in favor of the maximum. This is called a maximum likelihood decoder. The likelihood functions are given or computed from the specifications of the channel. Generally it is more convenient to compare the quantities $\log P(\mathbf{y} \mid \mathbf{x}^{(m)})$ called the log-likelihood functions and the result is unaltered since the logarithm is a monotonic function of its (always positive) argument.

To illustrate, let us consider again the BSC. Here each transmitted symbol is altered with probability $p < 1/2$. Now suppose we have received a particular N-dimensional binary sequence $\mathbf{y}$ and are considering a possible transmitted N-dimensional code sequence $\mathbf{x}^{(m)}$ which differs in d_m symbols from $\mathbf{y}$ (that is, the Hamming distance between $\mathbf{x}^{(m)}$ and $\mathbf{y}$ is d_m). Then since the channel is memoryless (i.e., it affects each symbol independently of all the others), the probability

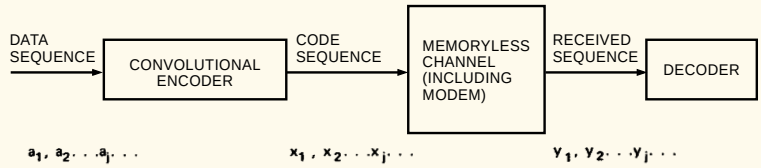


Fig. 10. Communication system employing convolutional codes.

that this $\mathbf{x}^{(m)}$ was transformed to the specific received $\mathbf{y}$ at distance d_m from it is

$$P(\mathbf{y} \mid \mathbf{x}^{(m)}) = p^{d_m}(1-p)^{N-d_m}$$

and the log-likelihood function is thus

$$\log P(\mathbf{y} \mid \mathbf{x}^{(m)}) = -d_m \log (1-p/p) + N \log (1-p)$$

Now if we compute this quantity for each possible transmitted sequence, it is clear that the second term is constant in each case. Furthermore, since we may assume $p < 1/2$ (otherwise the role of 0 and 1 is simply interchanged at the receiver), we may express this as

$$\log P(\mathbf{y} \mid \mathbf{x}^{(m)}) = -\alpha d_m - \beta \tag{3}$$

where α and β are positive constants and d_m is the (positive) distance. Consequently, it is clear that maximizing the log-likelihood function is equivalent to minimizing the Hamming distance d_m. Thus for the BSC to minimize the error probability we should choose that code sequence at minimum distance from the received sequence, as we have indicated and done in preceding sections.

We now consider a more physical practical channel: the AWGN channel with biphase[4] phase-shift keying (PSK) modulation. The modulator and optimum demodulator (correlator or integrate-and dump filter) for this channel are shown in Fig. 11.

We use the notation that x_{jk} is the kth code symbol for the jth branch. Each binary symbol (which we take here for convenience to be ± 1) modulates the carrier by $\pm\Pi/2$ radians for T seconds. The transmission rate is, therefore, $1/T$ symbols/second or $b/nT = R/T$ bit/s. The function ϵ_s is the energy transmitted for each symbol. The energy per bit is, therefore $\epsilon_b = \epsilon_s/R$. The white Gaussian noise is a zero-mean random process of one-sided spectral density N_0 W/Hz, which affects each symbol independently. It then follows directly that the channel output symbol y_{jk} is a Gaussian random variable whose mean is $\sqrt{\epsilon_s}x_{jk}$ (i.e., $+\sqrt{\epsilon_s}$ if $x_{jk} = 1$ and $-\sqrt{\epsilon_s}$ if $x_{jk} = -1$) and whose variance is $N_0/2$. Thus the conditional probability density (or likelihood) function of y_{jk} given x_{jk} is

$$p(y_{jk} \mid x_{jk}) = \frac{\exp\left[-(y_{jk} - \sqrt{\epsilon_s}\, x_{jk})^2/N_0\right]}{\sqrt{\Pi N_0}}. \tag{4}$$

The likelihood function for the jth branch of a particular

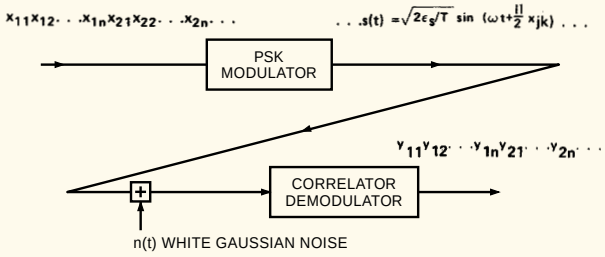


Fig. 11. Modem for additive white Gaussian noise PSK modulated memoryless channel.

code path $\mathbf{x}_j^{(m)}$

$$p(\mathbf{y}_i \mid \mathbf{x}_i^{(m)}) = \prod_{k=1}^{n} p(y_{ik} \mid x_{ik}^{(m)})$$

since each symbol is affected independently by the white Gaussian noise, and thus the log-likelihood function for the jth branch is

$$\begin{aligned}
\ln p(\mathbf{y}_i \mid \mathbf{x}_i^{(m)}) &= \sum_{k=1}^{n} \ln p(y_{ik} \mid x_{ik}^{(m)}) \\
&= -\frac{1}{N_0}\sum_{k=1}^{n}[y_{ik} - \sqrt{\epsilon_s}\, x_{ik}^{(m)}]^2 - \tfrac{1}{2}\ln\frac{\Pi}{N_0} \\
&= \frac{2\sqrt{\epsilon_s}}{N_0}\sum_{k=1}^{n} y_{ik}x_{ik}^{(m)} - \frac{\epsilon_s}{N_0}\sum_{k=1}^{n}[x_{ik}^{(m)}]^2 \\
&\quad - \frac{1}{N_0}\sum_{k=1}^{n} y_{ik}^2 - \tfrac{1}{2}\ln\frac{\Pi}{N_0} \\
&= C\sum_{k=1}^{n} y_{ik}x_{ik}^{(m)} - D
\end{aligned} \tag{5}$$

where C and D are independent of m, and we have used the fact that $[x_{jk}^{(m)}]^2 = 1$. Similarly, the log-likelihood[5] function for any path is the sum of the log-likelihood functions for each of its branches.

We have thus shown that the maximum likelihood decoder for the memoryless AWGN biphase (or quadriphase) modulated channel is one which forms the inner product between the received (real number) sequence and the code sequence (consisting of ± 1) and chooses the path corresponding to the greatest. Thus the *metric* for this channel is the inner product (5) as contrasted with the distance[6] metric used for the BSC.

[4] The results are the same for quadriphase PSK with coherent reception. The analysis proceeds in the same way, if we treat quadriphase PSK as two parallel independent biphase PSK channels.

[5] We have used the natural logarithm here, but obviously a change of base results merely in a scale factor.

[6] Actually it is easily shown that maximizing an inner product is equivalent to minimizing the Euclidean distance between the corresponding vectors.

For convolutional codes the structure of the code paths was described in Sections II–V. In Section III the optimum decoder was derived for the BSC. It now becomes clear that if we substitute the inner product metric $\Sigma y_{jk}x_{jk}^{(m)}$ for the distance metric $\Sigma d_{jk}^{(m)}$, used for the BSC, all the arguments used in Section III for the latter apply equally to this Gaussian channel. In particular the optimum decoder has a block diagram represented by the code state diagram. At step j the stored metric for each state (which is the maximum of the metrics of all the paths leading to this state at this time) is augmented by the branch metrics for branches emanating from this state. The comparisons are performed among all pairs of (or in general sets of 2^b) branches entering each state and the *maxima* are selected as the new most likely paths. The history (input data) of each new survivor must again be stored and the decoder is now ready for step $j + 1$.

Clearly, this argument generalizes to any memoryless channel and we must simply use the appropriate metric $\ln P(\mathbf{y} \mid \mathbf{x}^{(m)})$, which may always be determined from the statistical description of the channel. This includes, among others, AWGN channels employing other forms of modulation.[7]

In the next section, we apply the analysis of convolutional code distance properties of Section IV to determine the error probabilities of specific codes on more general memoryless channels.

VII. Performance of Convolutional Codes on Memoryless Channels

In Section IV we analyzed the distance properties of convolutional codes employing a state-diagram generating function technique. We now extend this approach to obtain tight upper bounds on the error probability of such codes. We shall consider the BSC, the AWGN channel and more general memoryless channels, in that order. We shall obtain both the first-event error probability, which is the probability that the correct path is excluded (not a survivor) for the first time at the jth step, and the bit error probability which is the expected ratio of bit errors to total number of bits transmitted.

A. Binary Symmetric Channel

The first-event error probability is readily obtained from the generating function $T(D)$ [(5) for the code of Fig. 1, which we shall again pursue for demonstrative purposes]. We may assume, without loss of generality, since we are dealing with group codes, that the all zeros path was transmitted. Then a first-event error is made at the jth step if this path is excluded by selecting another path merging with the all zeros at node a at the jth level.

Now suppose that the previous-level survivors were such that the path compared with the all zeros at step j is the path whose data sequence is $00 \cdots 0100$ corresponding to nodes $a \cdots a\ a\ b\ c\ a$ (see Fig. 4.). This differs from the correct (all zeros) path in five symbols. Consequently an error will be made in this comparison if the BSC caused three or more errors in these particular five symbols. Hence the probability of an error in this specific comparison is

$$P_5 = \sum_{e=3}^{5} \binom{5}{e} p^e (1-p)^{5-e}. \tag{6}$$

On the other hand, there is no assurance that this particular distance five path will have previously survived so as to be compared with the correct path at the jth step. If either of the distance 6 paths were compared instead, then four or more errors in the six different symbols will definitely cause an error in the survivor decision, while three errors will cause a tie which, if resolved by coin flipping, will result in an error only half the time. Then the probability if this comparison is made is

$$P_6 = \frac{1}{2}\binom{6}{3} p^3 (1-p)^3 + \sum_{e=4}^{6} \binom{6}{e} p^e (1-p)^{6-e}. \tag{7}$$

Similarly, if the previously surviving paths were such that a distance d path is compared with the correct path at the jth step, the resulting error probability is

$$P_k = \begin{cases} \displaystyle\sum_{e=(k+1)/2}^{k} \binom{k}{e} p^e (1-p)^{k-e}, & k \text{ odd} \\ \displaystyle\frac{1}{2}\binom{k}{k/2} p^{k/2}(1-p)^{k/2} + \sum_{e=k/2+1}^{k} \binom{k}{e} p^e (1-p)^{k-e}, & k \text{ even.} \end{cases} \tag{8}$$

Now at step j, since there is no simple way of determining previous survivors, we may overbound the probability of a first-event error by the sum of the error probabilities for all possible paths which merge with the correct path at this point. Note this *union bound* is indeed an upper bound because two or more such paths may both have distance closer to the received sequence than the correct path (even though only one has survived to this point) and thus the events are not disjoint. For the example with generating function (1) it follows that the first-event error probability[8] is bounded by

$$P_E < P_5 + 2P_6 + 4P_7 + \cdots + 2^k P_{k+5} + \cdots \tag{9}$$

where P_k is given by (8).

In Section VII-C it will be shown that (8) can be upper bounded by (see (39)).

$$P_k < 2^k p (1-p)^{k/2}. \tag{10}$$

Using this, the first-event error probability bound (9)

[7] Although more elaborate modulators, such as multiple FSK or multiphase modulators, might be employed, Jacobs [11] has shown that the most effective as well as the simplest system for wide-band space and satellite channels is the binary PSK modulator considered in the example of this section. We note again that the performance of quadriphase modulation is the same as for biphase modulation, when both are coherently demodulated.

[8] We are ignoring the finite length of the path, but the expression is still valid since it is an upper bound.

can be more loosely bounded by

$$P_E < \sum_{k=5}^{\infty} 2^{k-5} 2^k p(1-p)^{k/2}$$

$$= \frac{[2\sqrt{p(1-p)}]^5}{1 - 4\sqrt{p(1-p)}} = T(D)\,|_{D=2\sqrt{p(1-p)}} \qquad (11)$$

where $T(D)$ is just the generating function of (1)

It follows easily that for a general binary-tree ($b = 1$) convolutional code with generating function

$$T(D) = \sum_{k=d}^{\infty} a_k D^k \qquad (12)$$

the first-event error probability is bounded by the generalization of (9).

$$P_E < \sum_{k=d}^{\infty} a_k P_k \qquad (13)$$

where P_k is given by (8) and more loosely upper bounded by the generalization of (11)

$$P_E < T(D)\,|_{D=2\sqrt{p(1-p)}}. \qquad (14)$$

Whenever a decision error occurs, one or more bits will be incorrectly decoded. Specifically, those bits in which the path selected differs from the correct path will be incorrect. If only one error were ever made in decoding an arbitrary long code path, the number of bits in error in this incorrect path could easily be obtained from the augmented generating function $T(D, N)$ (such as given by (2) with factors in L deleted). For the exponents of the N factors indicate the number of bit errors for the given incorrect path arriving at node a at the jth level.

After the first error has been made, the incorrect paths no longer will be compared with a path which is overall correct, but rather with a path which has diverged from the correct path over some span of branches (see Fig. 12). If the correct path **x** has been excluded by a decision error at step j in favor of path **x′**, the decision at step $j + 1$ will be between **x′** and **x″**. Now the (first-event) error probability of (13) or (14) is for a comparison, at any step, between path **x** and any other path merging with it at that step, including path **x″** in this case. However, since the metric[9] for path **x′** is greater than the metric for **x**, for on this basis the correct path was excluded at step j, the probability that path **x″** metric exceeds path **x′** metric at step $j + 1$ is less than the probability that path **x″** exceeds the (correct) path **x** metric at this point. Consequently, the probability of a new incorrect path being selected after a previous error has occurred is upper bounded by the first-event error probability at that step.

Moreover, when a second error follows closely after a first error, it often occurs (as in Fig. 12) that the erroneous bit(s) of path **x″** overlap the erroneous bit(s) of path **x′**. With this in mind, we now show that for a

[9] Negative distance from the received sequence for the BSC, but clearly this argument generalizes to any memoryless channel.

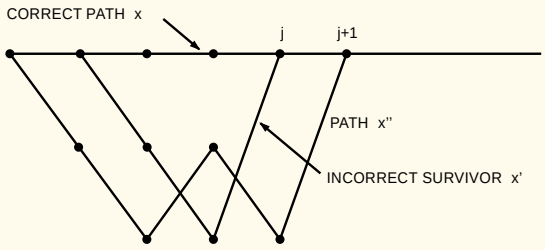


Fig. 12. Example of decoding decision after initial error has occurred.

binary-tree code if we weight each term of the first-event error probability bound at any step by the number of erroneous bits for each possible erroneous path merging with the correct path at that node level, we upper bound the bit error probability. For, a given step decision corresponds to decoder action on one more bit of the transmitted data sequence; the first-event error probability union bound with each term weighted by the corresponding number of bit errors is an upper bound on the expected number of bit errors caused by this action. Summing the expected number of bit errors over L steps, which as was just shown may result in overestimating through double counting, gives an upper bound on the expected number of bit errors in L branches for arbitrary L. But since the upper bound on expected number of bit errors is the same at each step, it follows, upon dividing the sum of L equal terms by L, that this expected number of bit errors per step is just the bit error probability P_B, for a binary-tree code ($b = 1$). If $b > 1$, then we must divide this expression by b, the number of bits encoded and decoded per step.

To illustrate the calculation of P_B for a convolutional code, let us consider again the example of Fig. 1. Its transfer function in D and N is obtained from (2), letting $L = 1$, since we are not now interested in the lengths of incorrect paths, to be

$$T(D, N) = \frac{D^5 N}{1 - 2DN}$$

$$= D^5 N + 2D^6 N^2 = \cdots + 2^k D^{k+5} N^{k+1} + \cdots . \qquad (15)$$

The exponents of the factors in N in each term determine the number of bit errors for the path(s) corresponding to that term. Since $T(D) = T(D, N)\,|_{N=1}$ yields the first-event error probability P_E, each of whose terms must be weighted by the exponent of N to obtain P_B, it follows that we should first differentiate $T(D, N)$ at $N = 1$ to obtain

$$\left.\frac{dT(D, N)}{dN}\right|_{N=1}$$

$$= D^5 + 2 \cdot 2D^6 + 3 \cdot 4D^7 + \cdots + (k+1)2^k D^{k+5} + \cdots$$

$$= \frac{D^5}{(1 - 2D)^2}. \qquad (16)$$

Then from this we obtain, as in (9), that for the BSC

$$P_B < P_5 + 2\cdot 2P_6 + 3\cdot 4P_7 + \cdots + (k+1)2^k P_{k+5} + \cdots \quad (17)$$

where P_k is given by (8).

If for P_k we use the upper bound (10) we obtain the weaker but simpler bound

$$P_B < \sum_{k=5}^{\infty} (k-4)2^{k-5}[4p(1-p)]^{k/2} = \frac{dT(D,N)}{dN}\bigg|_{N=1, D=2\sqrt{p(1-p)}} = \frac{|2\sqrt{p(1-p)}|^5}{[1-4\sqrt{p(1-p)}]^2}. \quad (18)$$

More generally for any binary-tree ($b = 1$) code used on the BSC if

$$\frac{dT(D,N)}{dN}\bigg|_{N=1} = \sum_{k=d}^{\infty} c_k D^k \quad (19)$$

then corresponding to (17)

$$P_B < \sum_{k=d}^{\infty} c_k P_k \quad (20)$$

and corresponding to (18) we have the weaker bound

$$P_B < \frac{dT(D,N)}{dN}\bigg|_{N=1, D=2\sqrt{p(1-p)}}. \quad (21)$$

For a nonbinary-tree code ($b \neq 1$), all these expressions must be divided by b.

The results of (14) and (18) will be extended to more general memoryless channels, but first we shall consider one more specific channel of particular interest.

B. AWGN Biphase-Modulated Channel

As was shown in Section VI the decoder for this channel operates in exactly the same way as for the BSC, except that instead of Hamming distance it uses the metric

$$\sum_i \sum_{j=1}^{n} x_{ij} y_{ij}$$

where $x_{ij} = \pm 1$ are the transmitted code symbols, y_{ij} the corresponding received (demodulated) symbols, and j runs over the n symbols of each branch while i runs over all the branches in a particular path. Hence, to analyze its performance we may proceed exactly as in Section VII-A except that the appropriate pairwise-decision errors P_k must be substituted for those of (6) to (8).

As before we assume, without loss of generality, that the correct (transmitted) path $\mathbf{x}$ has $x_{ij} = +1$ for all i and j (corresponding to the all zeros if the input symbols were 0 and 1). Let us consider an incorrect path $\mathbf{x}'$ merging with the correct path at a particular step, which has k negative symbols ($x_{ij}' = -1$) and the remainder positive. Such a path may be incorrectly chosen only if it has a higher metric than the correct path, i.e.,

$$\sum_i \sum_{j=1}^{n} x_{ij}' y_{ij} \geq \sum_i \sum_{j=1}^{n} x_{ij} y_{ij}$$

or

$$\sum_i \sum_{j=1}^{n} (x_{ij}' - x_{ij}) y_{ij} \geq 0$$

where i runs over all branches in the two paths. But since, as we have assumed, the paths $\mathbf{x}$ and $\mathbf{x}'$ differ in exactly k symbols, wherein $x_{ij} = 1$ and $x_{ij}' = -1$, the pairwise error probability is just

$$P_k = \Pr\left\{\sum_i \sum_{j=1}^{n} (x_{ij}' - x_{ij}) y_{ij} \geq 0\right\} = \Pr\left\{\sum_{r=1}^{k} (x_r' - x_r) y_r \geq 0\right\} = \Pr\left\{-2\sum_{r=1}^{k} y_r \geq 0\right\} = \Pr\left\{\sum_{r=1}^{k} y_r \leq 0\right\} \quad (22)$$

where r runs over the k symbols wherein the two paths differ. Now it was shown in Section VI that the y_{ij} are independent Gaussian random variables of variance $N_0/2$ and mean $\sqrt{\epsilon_s} x_{ij}$, where x_{ij} is the actually transmitted code symbol. Since we are assuming that the (correct) transmitted path has $x_{ij} = +1$ for all i and j, it follows that y_{ij} or y_r has mean $\sqrt{\epsilon_s}$ and variance $N_0/2$. Therefore, since the k variables y_r are independent and Gaussian, the sum $Z = \sum_{r=1}^{k} y_r$ is also Gaussian with mean $k\sqrt{\epsilon_s}$ and variance $kN_0/2$.

Consequently,

$$P_k = \Pr(Z < 0) = \int_{-\infty}^{0} \frac{\exp(-Z - k\sqrt{\epsilon_s})^2/kN_0}{\sqrt{\Pi k N_0}}\, dZ = \int_{\sqrt{2k\epsilon_s/N_0}}^{\infty} \left[\frac{\exp(-x^2/2)}{\sqrt{2\Pi}}\right] dx \triangleq \operatorname{erfc} \sqrt{\frac{2k\epsilon_s}{N_0}}. \quad (23)$$

We recall from Section VI that ϵ_s is the symbol energy, which is related to the bit energy by $\epsilon_s = R\epsilon_b$, where $R = b/n$. The bound on P_E then follows exactly as in Section VII-A and we obtain the same general bound as (13)

$$P_E < \sum_{k=d}^{\infty} a_k P_k \quad (24)$$

where a_k are the coefficients of

$$T'(D) = \sum_{k=d}^{\infty} a_k D^k \quad (25)$$

and where d is the minimum distance between any two paths in the code. We may simplify this procedure considerably while loosening the bound only slightly for this channel by observing that for $x \geq 0$, $y \geq 0$,

$$\operatorname{erfc} \sqrt{x+y} \leq \exp\left(\frac{-y}{2}\right) \operatorname{erfc} \sqrt{x}. \quad (26)$$

Consequently, for $k \geq d$, letting $l = k - d$, we have from (23)

$$P_k = \operatorname{erfc} \sqrt{\frac{2k\epsilon_s}{N_0}} = \operatorname{erfc} \sqrt{\frac{2(d+l)\epsilon_s}{N_0}}$$
$$\leq \exp\left(\frac{-l\epsilon_s}{N_0}\right) \operatorname{erfc} \sqrt{\frac{2d\epsilon_s}{N_0}} \tag{27}$$

whence the bound of (24), using (27), becomes

$$P_E < \sum_{k=d}^{\infty} a_k P_k \leq \operatorname{erfc} \sqrt{\frac{2d\epsilon_s}{N_0}} \sum_{k=d}^{\infty} a_k \exp\left[\frac{-(k-d)\epsilon_s}{N_0}\right]$$

or

$$P_E < \operatorname{erfc} \sqrt{\frac{2d\epsilon_s}{N_0}} \exp\left(\frac{d\epsilon_s}{N_0}\right) T(D) \,|_{D=\exp(-\epsilon_s/N_0)}. \tag{28}$$

The bit error probability can be obtained in exactly the same way. Just as for the BSC [(19) and (20)] we have that for a binary-tree code

$$P_B < \sum_{k=d}^{\infty} c_k P_k \tag{29}$$

where c_k are the coefficients of

$$\left.\frac{dT(D, N)}{dN}\right|_{N=1} = \sum_{k=d}^{\infty} c_k D^k. \tag{30}$$

Thus following the came arguments which led from (24) to (28) we have for a binary-tree code

$$P_B < \operatorname{erfc} \sqrt{\frac{2d\epsilon_s}{N_0}} \exp\left(\frac{d\epsilon_s}{N_0}\right) \left.\frac{dT(D, N)}{dN}\right|_{N=1, D=\exp(-\epsilon_s/N_0)} \tag{31}$$

For $b > 1$, this expression must be divided by b.

To illustrate the application of this result we consider the code of Fig. 1 with parameters $K = 3$, $R = 1/2$, whose transfer function is given by (15). For this case since $R = 1/2$ and $\epsilon_s = 1/2\ \epsilon_b$, we obtain

$$P_B < \frac{\operatorname{erfc} \sqrt{5\epsilon_b/N_0}}{(1 - 2e^{-\epsilon_b/2N_0}}. \tag{32}$$

Since the number of states in the state diagram grows exponentially with K, direct calculation of the generating function becomes unmanageable for $K > 4$. On the other hand, a generating function calculation is basically just a matrix inversion (see Appendix I), which can be performed numerically for a given value of D. The derivative at $N = 1$ can be upper bounded by evaluating the first difference $[T(D, 1 + \epsilon) - T(D, 1)]/\epsilon$, for small ϵ. A computer program has been written to evaluate (31) for any constraint length up to $K = 10$ and all rates $R = 1/n$ as well as $R = 2/3$ and $R = 3/4$. Extensive results of these calculations are given in the paper by Heller and Jacobs [24], along with the results of simulations of the corresponding codes and channels. The simulations verify the tightness of the bounds.

In the next section, these bounding techniques will be extended to more general memoryless channels, from which (28) and (31) can be obtained directly, but without the first two factors. Since the product of the first two factors is always less than one, the more general bound is somewhat weaker.

C. General Memoryless Channels

As was indicated in Section VI, for equally likely input data sequences, the minimum error probability decoder chooses the path which maximizes the log-likelihood function (metric)

$$\ln P(\mathbf{y} \mid \mathbf{x}^{(m)})$$

over all possible paths $\mathbf{x}^{(m)}$. If each symbol is transmitted (or modulates the transmitter) independent of all preceding and succeeding symbols, and the interference corrupts each symbol independently of all the others, then the channel, which includes the modem, is said to be memoryless[10] and the log-likelihood function

$$\ln P(\mathbf{y} \mid \mathbf{x}^{(m)}) = \sum_i \sum_{j=1}^{n} \ln P(y_{ij} \mid x_{ij}^{(m)})$$

where $x_{ij}^{(m)}$ is a code symbol of the mth path, y_{ij} is the corresponding received (demodulated) symbol, j runs over the n symbols of each branch, and i runs over the branches in the given path. This includes the special cases considered in Sections VII-A and -B.

The decoder is the same as for the BSC except for using this more general metric. Decisions are made after each set of new branch metrics have been added to the previously stored metrics. To analyze performance, we must merely evaluate P_k, the pairwise error probability for an incorrect path which differs in k symbols from the correct path, as was done for the special channels of Sections VII-A and -B. Proceeding as in (22), letting x_{ij} and x_{ij}' denote symbols of the correct and incorrect paths, respectively, we obtain

$$\begin{aligned} P_k(\mathbf{x}, \mathbf{x}') &= \Pr\left[\sum_i \sum_{j=1}^{n} \ln P(y_{ij} \mid x_{ij}') > \sum_i \sum_{j=1}^{n} \ln P(y_{ij} \mid x_{ij})\right] \\ &= \Pr\left\{\sum_{r=1}^{k} \ln \frac{P(y_r \mid x_r')}{P(y_r \mid x_r)} > 0\right\} \\ &= \Pr\left\{\prod_{r=1}^{k} \frac{P(y_r \mid x_r')}{P(y_r \mid x_r)} > 1\right\} \end{aligned} \tag{33}$$

where r runs over the k code symbols in which the paths differ. This probability can be rewritten as

$$P_k(\mathbf{x}, \mathbf{x}') = \sum_{\mathbf{y} \in Y_k} \prod_{r=1}^{k} P(y_r \mid x_r) \tag{34}$$

where Y_k is the set of all vectors $\mathbf{y} = (y_1, y_2, \cdots, y_r, \cdots, y_k)$ for which

[10] Often more than one code symbol in a given branch is used to modulate the transmitter at one time. In this case, provided the interference still affects succeeding branches independently, the channel can still be treated as memoryless but now the symbol likelihood functions are replaced by branch likelihood functions and (33) is replaced by a single sum over i.

$$\prod_{r=1}^{k} \frac{P(y_r \mid x_r')}{P(y_r \mid x_r)} > 1. \tag{35}$$

But if this is the case, then

$$P_k(\mathbf{x}, \mathbf{x}') < \sum_{\mathbf{y} \in Y_k} \prod_{r=1}^{k} P(y_r \mid x_r) \left[\frac{P(y_r \mid x_r')}{P(y_r \mid x_r)}\right]^{1/2}$$

$$< \sum_{\text{all } \mathbf{y} \in Y} \prod_{r=1}^{k} P(y_r \mid x_r)^{1/2} P(y_r \mid x_r')^{1/2} \tag{36}$$

where Y is the entire space of received vectors.[11] The first inequality is valid because we are multiplying the summand by a quantity greater than unity,[12] and the second because we are merely extending the sum of positive terms over a larger set. Finally we may break up the k-dimensional sum over $\mathbf{y}$ into k one-dimensional summations over $y_1, y_2, \cdots, y_k$, respectively, and this yields

$$P_k(\mathbf{x}, \mathbf{x}') \leq \sum_{y_1} \sum_{y_2} \cdots \sum_{y_k} \prod_{r=1}^{k} P(y_r \mid x_r)^{1/2} P(y_r \mid x_r')^{1/2}$$

$$= \prod_{r=1}^{k} \sum_{y_r} P(y_r \mid x_r)^{1/2} P(y_r \mid x_r')^{1/2} \tag{37}$$

To illustrate the use of this bound we consider the two specific channels treated above. For the BSC, y_r is either equal to x_r, the transmitted symbol, or to $\bar{x}_r$, its complement. Now y_r depends on x_r through the channel statistics. Thus

$$P(y_r = x_r) = 1 - p$$

$$P(y_r = \bar{x}_r) = p. \tag{38}$$

For each symbol in the set $r = 1, 2, \cdots, k$ by definition $x_r \neq x_r'$. Hence for each term in the sum if $x_r = 0$, $x_r' = 1$ or vice versa. Hence, whatever x_r and x_r' may be

$$\sum_{y_r=0}^{1} P(y_r \mid x_r)^{1/2} P(y_r \mid x_r')^{1/2} = 2p^{1/2}(1-p)^{1/2}$$

and the product (37) of k identical factors is

$$P_k = 2^k p^{k/2} (1-p)^{k/2} \tag{39}$$

for all pairs of correct and incorrect paths. This was used in Section VII-A to obtain the bounds (11) and (21).

For the AWGN channel of Section VII-B we showed that the likelihood functions (probability densities) were

$$p(y_r \mid x_r) = \frac{\exp\left[-(y_r - \sqrt{\epsilon_s}\, x_r)^2 / N_0\right]}{\sqrt{\pi N_0}} \tag{40}$$

where $x_r = +1$ or -1 and

$$x_r + x_r' = 0. \tag{41}$$

Since y_r is a real variable, the space of y_r is the real line and the sum in (37) becomes the integral

$$\int_{-\infty}^{\infty} p(y_r \mid x_r)^{1/2} p(y_r \mid x_r')^{1/2}\, dy_r = \frac{1}{\sqrt{\pi N_0}} \int_{-\infty}^{\infty} \exp\left\{\frac{-[(y_r - \sqrt{\epsilon_s}\, x_r)^2 + (y_r - \sqrt{\epsilon_s}\, x_r')^2]}{2N_0}\right\} dy_r$$

$$= \frac{1}{\sqrt{\pi N_0}} \int_{-\infty}^{\infty} \exp\left[\frac{-(y_r^2 + \epsilon_s)}{N_0}\right] dy_r = \exp\left(\frac{-\epsilon_s}{N_0}\right)$$

where we have used (41) and $x_r^2 = x_r'^2 = 1$. The product of these k identical terms is, therefore,

$$P_k < \exp\left(\frac{-k\epsilon_s}{N_0}\right) \tag{42}$$

for all pairs of correct and incorrect paths. Inserting these bounds in the general expressions (24) and (29), and using (25) and (30) yields the bound on first-event error probability and bit error probability.

$$P_E < T(D)\,|_{D = \exp(-\epsilon_s/N_0)} \tag{43}$$

$$P_B < \left.\frac{dT(D, N)}{dN}\right|_{N=1, D=\exp(-\epsilon_s/N_0)} \tag{44}$$

which are somewhat (though not exponentially) weaker than (28) and (31).

A characteristic feature of both the BSC and the AWGN channel is that they affect each symbol in the same way independent of its location in the sequence. Any memoryless channel has this property provided it is stationary (statistically time invariant). For a stationary memoryless channel (37) reduces to

$$P_k(\mathbf{x}, \mathbf{x}') < \left[\sum_{y_r} P(y_r \mid x_r)^{1/2} P(y_r \mid x_r')^{1/2}\right]^k \triangleq D_0^k \tag{45}$$

where[13]

$$D_0 \triangleq \sum_{y_r} P(y_r \mid x_r)^{1/2} P(y_r \mid x_r')^{1/2} < 1. \tag{46}$$

While this bound on P_k is valid for all such channels, clearly it depends on the actual values assumed by the symbols x_r and x_r', of the correct and incorrect path, and these will generally vary according to the pairs of paths $\mathbf{x}$ and $\mathbf{x}'$ in question. However, if the input symbols are binary, x and $\bar{x}$, whenever $x_r = x$, then $x_r' = \bar{x}$,

[11] This would be the set of all 2^k k-dimensional binary vectors for the BSC, and Euclidean k space for the AWGN channel. Note also that the bound of (36) may be improved for asymmetric channels by changing the two exponents of ½ to s and $1 - s$, respectively, where $0 < s < 1$.

[12] The square root of a quantity greater than one is also greater than one.

[13] For an asymmetric channel this bound may be improved by changing the two exponents 1/2 to s and $1 - s$, respectively, where $0 < s < 1$.

so that for any input-binary memoryless channel (46) becomes

$$D_0 = \sum_y P(y \mid x)^{1/2} P(y \mid \bar{x})^{1/2} \tag{47}$$

and consequently

$$P_E < T(D) \mid_{D=D_0} \tag{48}$$

$$P_B < \frac{dT(D, N)}{dN} \bigg|_{N=1, D=D_0} \tag{49}$$

where D_0 is given by (47). Other examples of channels of this type are FSK modulation over the AWGN (both coherent and noncoherent) and Rayleigh fading channels.

VIII. Systematic and Nonsystematic Convolutional Codes

The term *systematic* convolutional code refers to a code on each of whose branches one of the code symbols is just the data bit generating that branch. Thus a systematic coder will have its stages connected to only $n - 1$ adders, the nth being replaced by a direct line from the first stage to the commutator. Fig. 13 shows an $R = 1/2$ systematic coder for $K = 3$.

It is well known that for group block codes, any nonsystematic code can be transformed into a systematic code which performs exactly as well. This is not the case for convolutional codes. The reason for this is that, as was shown in Section VII, the performance of a code on any channel depends largely on the relative distances between codewords and particularly on the minimum free distance d, which is the exponent of D in the leading term of the generating function. Eliminating one of the adders results in a reduction of d. For example, the maximum free distance code for $K = 3$ is that of Fig. 13 and this has $d = 4$, while the nonsystematic $K = 3$ code of Fig. 1 has minimum free distance $d = 5$. Table I shows the maximum minimum free distance for systematic and nonsystematic codes for $K = 2$ through 5. For large constraint lengths the results are even more widely separated. In fact, Bucher and Heller [19] have shown that for asymptotically large K, the performance of a systematic code of constraint length K is approximately the same as that of a nonsystematic code of constraint length $K(1 - R)$. Thus for $R = 1/2$ and very large K, systematic codes have the performance of nonsystematic codes of half the constraint length, while requiring exactly the same optimal decoder complexity. For $R = 3/4$, the constraint length is effectively divided by 4.

IX. Catastrophic Error Propagation in Convolutional Codes

Massey and Sain [13] have defined a catastrophic error as the event that a finite number of channel symbol errors causes an infinite number of data bit errors to be decoded. Furthermore, they showed that a necessary and sufficient condition for a convolutional code to produce

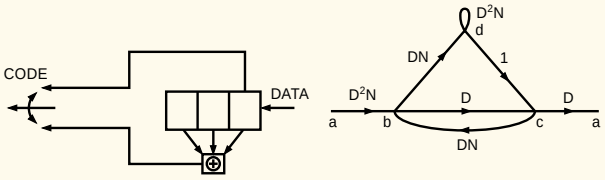


Fig. 13. Systematic convolutional coder for $K = 3$, and $r = 1/2$.

TABLE I
Maximum–Minimum Free Distance

K	Systematic	Nonsystematic[a]
2	3	3
3	4	5
4	4	6
5	5	7

[a] We have excluded catastrophic codes (see Section IX); $R = \frac{1}{2}$.

catastrophic errors is that all of the adders have tap sequences, represented as polynomials, with a common factor.

In terms of the state diagram it is easily seen that catastrophic errors can occur if and only if any closed loop path in the diagram has a zero weight (i.e, the exponent of D for the loop path is zero). To illustrate this, we consider the example of Fig. 14.

Assuming that the all zeros is the correct path, the incorrect path $a\ b\ d\ d\ \cdots\ d\ c\ a$ has exactly 6 ones, no matter how many times we go around the self loop d. Thus for a BSC, for example, four-channel errors may cause us to choose this incorrect path or consequently make an arbitrarily large number of bit errors (equal to two plus the number of times the self loop is traversed). Similarly for the AWGN channel this incorrect path with arbitrarily many corresponding bit errors will be chosen with probability erfc $\sqrt{6\epsilon_b/N_0}$.

Another necessary and sufficient condition for catastrophic error propagation, recently found by Odenwalder [20] is that any nonzero data path in the trellis or state diagram produces $K - 1$ consecutive branches with all zero code symbols.

We observe also that for binary-tree ($R = 1/n$) codes, if each adder of the coder has an even number of connections, then the self loop corresponding to the all ones (data) state will have zero weight and consequently the code will be catastrophic.

The main advantage of a systematic code is that it can never be catastrophic, since each closed loop must contain at least one branch generated by a nonzero data bit and thus having a nonzero code symbol. Still it can be shown [23] that only a small fraction of nonsystematic codes is catastrophic (in fact, $1/(2^n - 1)$ for binary-tree $R = 1/n$ codes. We note further that if catastrophic errors are ignored, nonsystematic codes with even larger free distance than those of Table I exist.

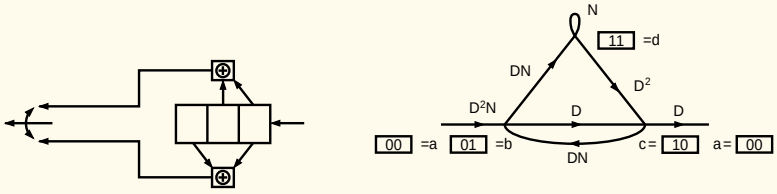


Fig. 14. Coder displaying catastrophic error propagation.

X. Performance Bounds for Best Convolutional Codes for General Memoryless Channels and Comparison with Block Codes

We begin by considering the path structure of a binary-tree[14] ($b = 1$) convolutional code of any constraint K, independent of the specific coder used. For this purpose we need only determine $T(L)$ the generating function for the state diagram with each branch labeled merely by L so that the exponent of each term of the infinite series expansion of $T(L)$ determines the length over which an incorrect path differs from the correct path before merging with it at a given node level. (See Fig. 7 and (2) with $D = N = 1$).

After some manipulation of the state-transition matrix of the state diagram of a binary-tree convolutional code of constraint length K, it is shown in Appendix I[15] that

$$T(L) = \frac{L^K(1-L)}{1-2L+L^K} < \frac{L^K}{1-2L}$$

$$= L^K(1 + 2L + 4L^2 + \cdots + 2^kL^k + \cdots) \qquad (50)$$

where the inequality indicates that more paths are being counted than actually exist. The expression (50) indicates that of the paths merging with the correct path at a given node level there is no more than one of length K, no more than two of length $K + 1$, no more than three of length $K + 2$, etc.

We have purposely avoided considering the actual code or coder configuration so that the preceding expressions are valid for all binary-tree codes of constraint length K. We now extend our class of codes to include time-varying convolutional codes. A time-varying coder is one in which the tap positions may be changed after each shift of the bits in the register. We consider the ensemble of all possible time-varying codes, which includes as a subset the ensemble of all fixed codes, for a given constraint length K. We further impose a uniform probabilistic measure on all codes in this ensemble by randomly reselecting each tap position after each shift of the register. This can be done by hypothetically flipping a coin nK times after each shift, once for each stage of the register and for each of the n adders. If the outcome is a head we connect the particular stage to the particular adder; if it is a tail we do not. Since this is repeated for each new branch, the result is that for each branch of the trellis the code sequence is a random binary n-dimensional vector. Furthermore, it can be shown that the distribution of these random code sequences is the same for each branch at each node level except for the all zeros path, which must necessarily produce the all zeros code sequence on each branch. To avoid treating the all zeros path differently, we ensure statistical uniformity by requiring further that after each shift a random binary n-dimensional vector be added to each branch[16] and that this also be reselected after each shift. (This additional artificiality is unnecessary for input-binary channels but is required to prove our result for general memoryless channels). Further details of this procedure are given in Viterbi [9].

We now seek a bound on the average error probability of this ensemble of codes relative to the measure (random-selection process) imposed. We begin by considering the probability that after transmission over a memoryless channel the metric of one of the fewer than 2^k paths merging with the correct path after differing in $K + k$ branches, is greater than the correct metric. Let $\mathbf{x}_i$ be the correct (transmitted) sequence and $\mathbf{x}_i'$ an incorrect sequence for the ith branch of the two paths. Then following the argument which led to (37) we have that the probability that the given incorrect path may cause an error is bounded by

$$P_{K+k}(\mathbf{x}, \mathbf{x}') < \prod_{i=1}^{K+k} \sum_{\mathbf{y}_i} P(\mathbf{y}_i \mid \mathbf{x}_i)^{1/2} P(\mathbf{y}_i \mid \mathbf{x}_i')^{1/2} \qquad (51)$$

where the product is over all $K + k$ branches in the path. If we now average over the ensemble of codes constructed above we obtain

$$\bar{P}_{K+k} < \prod_{i=1}^{K+k} \sum_{\mathbf{x}_i} \sum_{\mathbf{x}_i'} \sum_{\mathbf{y}_i} q(\mathbf{x}_i) P(\mathbf{y}_i \mid \mathbf{x}_i)^{1/2} q(\mathbf{x}_i') P(\mathbf{y}_i \mid \mathbf{x}_i')^{1/2} \qquad (52)$$

where $q(\mathbf{x})$ is the measure imposed on the code symbols of each branch by the random selection, and because of the statistical uniformity of all branches we have

$$\bar{P}_{K+k} < \{\sum_{\mathbf{y}} [\sum_{\mathbf{x}} q(\mathbf{x}) P(\mathbf{y} \mid \mathbf{x})^{1/2}]^2\}^{K+k} = 2^{-(K+k)nR_0} \qquad (53)$$

[14] Although for clarity all results will be derived for $b = 1$, the extension to $b > 1$ is direct and the results will be indicated at the end of this Section.

[15] This generating function can also be used to obtain error bounds for orthogonal convolutional codes all of whose branches have the same weight, as is shown in Appendix I.

[16] The same vector is added to all branches at a given node level.

where

$$R_0 \triangleq -\frac{1}{n}\log_2\{\sum_{\mathbf{y}}[\sum_{\mathbf{x}} q(\mathbf{x})P(\mathbf{y}\mid\mathbf{x})^{1/2}]^2\}. \quad (54)$$

Note that the random vectors $\mathbf{x}$ and $\mathbf{y}$ are n dimensional. If each symbol is transmitted independently on a memoryless channel, such as was the case in the channels of Sections VII-A and -B, (54) is reduced further to

$$R_0 = -\log_2\{\sum_{y}[\sum_{x} q(x)P(y\mid x)^{1/2}]^2\} \quad (55)$$

where x and y are now scalar random variables associated with each code symbol. Note also that because of the statistical uniformity of the code, the results are independent of which path was transmitted and which incorrect path we are considering.

Proceeding as in Section VII, it follows that a union bound on the ensemble average of the first-event error probability is obtained by substituting $\bar{P}_{K+k}$ for L^{K+k} in (50). Thus

$$\bar{P}_E < \sum_{k=0}^{\infty} 2^k\bar{P}_{K+k} < \sum_{k=0}^{\infty} 2^k 2^{-(K+k)R_0/R} = \frac{2^{-KR_0/R}}{1-2^{-(R_0/R-1)}} \quad (56)$$

where we have used the fact that since $b = 1$, $R = 1/n$ bits/symbol.

To bound the bit error probability we must weight each term of (56) by the number of bit errors for the corresponding incorrect path. This could be done by evaluating the transfer function $T(L, N)$ as in Section VII (see also Appendix I), but a simpler approach, which yields a simpler bound which is nearly as tight, is to recognize that an incorrectly chosen path which merges with the correct path after $K + k$ branches can produce no more $k + 1$ bit errors. For, any path which merges with the correct path at a given level must be generated by data which coincides with the correct path data over the last $K - 1$ branches prior to merging, since only in this way can the coder register be filled with the same bits as the correct path, which is the condition for merging. Hence the number of incorrect bits due to a path which differs from the correct path in $K + k$ branches can be no greater than $K + k - (K - 1) = k + 1$.

Hence we may overbound $\bar{P}_B$ by weighting the kth term of (56) by $k + 1$, which results in

$$\bar{P}_B < \sum_{k=0}^{\infty}(k+1)2^{-k(R_0/R-1)}2^{-KR_0/R} = \frac{2^{-KR_0/R}}{[1-2^{-(R_0/R-1)}]^2}. \quad (57)$$

The bounds of (56) and (57) are finite only for rates $R < R_0$, and R_0 can be shown to be always less than the channel capacity.

To improve on these bounds when $R > R_0$, we must improve on the union bound approach by obtaining a single bound on the probability that any one of the fewer than 2^k paths which differ from the correct path in $K + k$ branches has a metric higher than the correct path at a given node level. This bound, first derived by Gallager [5] for block codes, is always less than 2^k times the bound for each individual path. Letting $Q_{K+k} \triangleq$ Pr (any one of 2^k incorrect path metrics $>$ correct path metric), Gallager [5] has shown that its ensemble average for the code ensemble is bounded by

$$\bar{Q}_{K+k} < 2^{k\rho}2^{-(K+k)nE_0(\rho)} \quad (58)$$

where

$$E_0(\rho) = -\frac{1}{n}\log_2\sum_{\mathbf{y}}[\sum_{\mathbf{x}} q(\mathbf{x})p(\mathbf{y}\mid\mathbf{x})^{1/(1+\rho)}]^{1+\rho}, \quad 0 < \rho \le 1 \quad (59)$$

where ρ is an arbitrary parameter which we shall choose to minimize the bound. It is easily seen that $E_0(0) = 0$, while $E_0(1) = R_0$, in which case $\bar{Q}_{K+k} = 2^k\bar{P}_{K+k}$, the ordinary union bound of (56). We bound the overall ensemble first-event error probability by the probability of the union of these composite events given by (58). Thus we find

$$\bar{P}_E < \sum_{k=0}^{\infty}\bar{Q}_{K+k} < \frac{2^{-KE_0(\rho)/R}}{1-2^{-(E_0(\rho)/R-\rho)}}. \quad (60)$$

Clearly (60) reduces to (56) when $\rho = 1$.

To determine the bit error probability using this approach, we must recognize that $\bar{Q}_{K+k}$ refers to 2^k different incorrect paths, each with a different number of incorrect bits. However, just as was observed in deriving (57), an incorrect path which differs from the correct path in $K + k$ branches prior to merging can produce at most $k + 1$ bit errors. Hence weighting the kth term of (60) by $k + 1$, we obtain

$$\bar{P}_B < \sum_{k=0}^{\infty}(k+1)\bar{Q}_{K+k} < \sum_{k=0}^{\infty}(k+1)2^{-k(E_0(\rho)/R-\rho)}2^{-KE_0(\rho)/R} = \frac{2^{-KE_0(\rho)/R}}{[1-2^{-(E_0(\rho)/R-\rho)}]^2}, \quad 0 < \rho \le 1. \quad (61)$$

Clearly (61) reduces to (57) when $\rho = 1$.

Before we can interpret the results of (56), (57), (60), and (61) it is essential that we establish some of the properties of $E_0(\rho)$ $(0 < \rho \le 1)$ defined by (59). It can be shown [5], [14] that for any memoryless channel, $E_0(\rho)$ is a concave monotonic nondecreasing function as shown in Fig. 15 with $E_0(0) = 0$ and $E_0(1) = R_0$.

Where the derivative $E_0'(\rho)$ exists, it decreases with ρ and it follows easily from the definition that

$$\lim_{\rho\to 0}E_0'(\rho) = \frac{1}{n}\sum_{\mathbf{y}}\sum_{\mathbf{x}} q(\mathbf{x})\log_2\frac{P(\mathbf{y}\mid\mathbf{x})}{\sum_{\mathbf{x}'}q(\mathbf{x}')P(\mathbf{y}\mid\mathbf{x}')} = \frac{1}{n}I(X^n, Y^n) \triangleq C \quad (62)$$

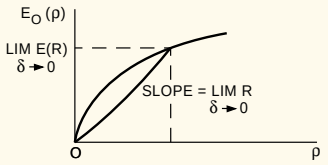


Fig. 15. Example of $E_0(\rho)$ function for general memoryless channel.

the mutual information of the channel[17] where X^n and Y^n are the channel input and output spaces, respectively, for each branch sequence. Consequently, it follows that to minimize the bounds (60) and (61), we must make $\rho \leq 1$ as large as possible to maximize the exponent of the numerator, but at the same time we must ensure that

$$R < \frac{E_0(\rho)}{\rho}$$

in order to keep the denominator positive. Thus since $E_0(1) = R_0$ and $E_0(\rho) < R_0$, for $\rho < 1$, it follows that for $R < R_0$ and sufficiently large K we should choose $\rho = 1$, or equivalently use the bounds (56) and (57). We may thus combine all the above bounds into the expressions

$$\bar{P}_E < \frac{2^{-KE(R)/R}}{1 - 2^{-\delta(R)}} \tag{63}$$

$$\bar{P}_B < \frac{2^{-KE(R)/R}}{[1 - 2^{-\delta(R)}]^2} \tag{64}$$

where

$$E(R) = \begin{cases} R_0, & 0 \leq R < R_0 \\ E_0(\rho), & R_0 < R < C, \quad 0 < \rho \leq 1 \end{cases} \tag{65}$$

$$\delta(R) = \begin{cases} R_0/R - 1, & 0 < R < R_0 \\ E_0(\rho)/R - \rho, & R_0 \leq R < C, \quad 0 < \rho \leq 1. \end{cases} \tag{66}$$

To minimize the numerators of (63) and (64) for $R > R_0$ we should choose ρ as large as possible, since $E_0(\rho)$ is a nondecreasing function of ρ. However, we are limited by the necessity of making $\delta(R) > 0$ to keep the denominator from becoming zero. On the other hand, as the constraint length K becomes very large we may choose $\delta(R) = \delta$ very small. In particular, as δ approaches 0, (65) approaches

$$\lim_{\delta \to 0} E(R) = \begin{cases} R_0, & 0 < R < R_0 \\ E_0(\rho), & R_0 \leq R = E_0(\rho)/\rho < C, \quad 0 < \rho \leq 1. \end{cases} \tag{67}$$

[17] C can be made equal to the channel capacity by properly choosing the ensemble measure $q(\mathbf{x})$. For an input-binary channel the random binary convolutional coder described above achieves this. Otherwise a further transformation of the branch sequence into a smaller set of nonbinary sequences is required [9].

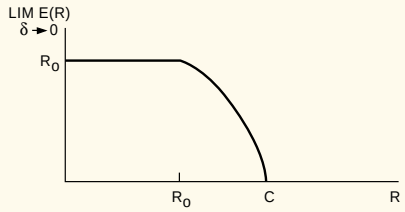


Fig. 16. Typical limiting value of exponent of (67).

Fig. 15 demonstrates the graphical determination of $\lim_{\delta \to 0} E(R)$ from $E_0(\rho)$.

It follows from the properties of $E_0(\rho)$ described, that for $R > R_0$, $\lim_{\delta \to 0} E(R)$ decreases from R_0 to 0 as R increases from R_0 to C, but that it remains positive for all rates less than C. The function is shown for a typical channel in Fig. 16.

It is particularly instructive to obtain specific bounds, in the limiting case, for the class of "very noisy" channels, which includes the BSC with $p = 1/2 - \gamma$ where $|\gamma| \ll 1$ and the biphase modulated AWGN with $\epsilon_s/N_0 \ll 1$. For this class of channels it can be shown [5] that

$$E_0(\rho) = \frac{\rho C}{1 + \rho} \tag{68}$$

and consequently $R_0 = E_0(1) = C/2$. (For the BSC, $C = \gamma^2/2 \ln 2$ while for the AWGN, $C = \epsilon_s/N_0 \ln 2$.)

For the very noisy channel, suppose we let $\rho = C/R - 1$, so that using (68) we obtain $E_0(\rho) = C - R$. Then in the limit as $\delta \to 0$ (65) becomes for a very noisy channel

$$\lim_{\delta \to 0} E(R) = \begin{cases} C/2, & 0 \leq R \leq C/2 \\ C - R, & C/2 \leq R \leq C. \end{cases} \tag{69}$$

This limiting form of $E(R)$ is shown in Fig. 17.

The bounds (63) and (64) are for the average error probabilities of the ensemble of codes relative to the measure induced by random selection of the time-varying coder tap sequences. At least one code in the ensemble must perform better than the average. Thus the bounds (63) and (64) hold for the best time-varying binary-tree convolutional coder of constraint length K. Whether there exists a fixed convolutional code with this performance is an unsolved problem. However, for small K the results of Section VII seem to indicate that these bounds are valid also for fixed codes.

To determine the tightness of the upper bounds, it is useful to have lower bounds for convolutional code error probabilities. It can be shown [9] that for all $R < C$

$$P_B \geq P_E > 2^{-K[E_L(R)/R - o(K)]} \tag{70}$$

where

$$\begin{aligned} E_L(R) &= E_0(\rho) \\ R &= E_0(\rho)/\rho \end{aligned}, \quad 0 \leq \rho < \infty, \quad 0 \leq R \leq C \tag{71}$$

and $o(K) \to 0$ as $K \to \infty$. Comparison of the parametric

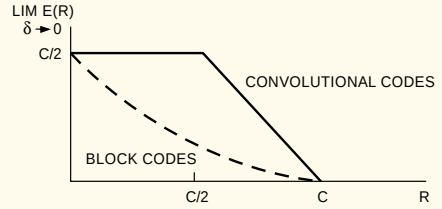


Fig. 17. Limiting values of $E(R)$ for very noisy channels.

equations (67) with (71), shows that

$$E_L(R) = \lim_{\delta \to 0} E(R)$$

for $R > R_0$ but is greater for low rates.

For very noisy channels, it follows easily from (71) and (68) that

$$E_L(R) = C - R, \qquad 0 \leq R \leq C.$$

Actually, however, tighter lower bounds for $R < C/2$ (Viterbi [9]) show that for very noisy channels

$$E_L(R) = \begin{cases} C/2, & 0 \leq R \leq C/2 \\ C - R, & C/2 \leq R < C, \end{cases} \tag{72}$$

which is precisely the result of (69) or of Fig. 17. It follows that, at least for very noisy channels, the exponential bounds are asymptotically exact.

All the results derived in this section can be extended directly to nonbinary ($b > 1$) codes. It is easily shown (Viterbi [9]) that the same results hold with $R = b/n$, R_0 and $E_0(\rho)$ multiplied by b, and all event probability upper bounds multiplied by $2^b - 1$, and bit probability upper bounds multiplied by $(2^b - 1)/b$.

Clearly, the ensemble of codes considered here is nonsystematic. However, by a modification of the arguments used here, Bucher and Heller [19] restricted the ensemble to systematic time-varying convolutional codes (i.e., codes for which b code symbols of each branch correspond to the data which generates the branch) and obtained all the above results modified only to the extent that the exponents $E(R)$ and $E_L(R)$ are multiplied by $1 - R$. (See also Section VIII.)

Finally, it is most revealing to compare the asymptotic results for the best convolutional codes of a given constraint length with the corresponding asymptotic results for the best block codes of a given block length. Suppose that K bits are coded into a block code of length N so that $R = K/N$ bits/code symbol. Then it can be shown (Gallager [5], Shannon *et al.* [8]) that for the best block code, the bit error probability is bounded above and below by

$$2^{-K[EL_b(R)/R + o(K)]} < P_B < 2^{-KE_b(R)/R} \tag{73}$$

where

$$E_b(R) = \max_{0 \leq \rho \leq 1} [E_0(\rho) - \rho R]$$

$$E_{Lb}(R) \leq \max_{0 \leq \rho} [E_0(\rho) - \rho R].$$

Both $E_b(R)$ and $E_{Lb}(R)$ are functions of R which for all $R > 0$ are less than the exponents $E(R)$ and $E_L(R)$ for convolutional codes [9]. In particular, for very noisy channels they both become [5]

$$E_b(R) = E_{Lb}(R) = \begin{cases} C/2 - R \\ (\sqrt{C} - \sqrt{R})^2. \end{cases} \tag{74}$$

This is plotted as a dotted curve in Fig. 17.

Thus it is clear by comparing the magnitudes of the negative exponents of (73) and (64) that, at least for very noisy channels, a convolutional code performs much better asymptotically than the corresponding block code of the same order of complexity. In particular at $R = C/2$, the ratio of exponents is 5.8, indicating that to achieve equivalent performance asymptotically the block length must be over five times the constraint length of the convolutional code. Similar degrees of relative performance can be shown for more general memoryless channels [9].

More significant from a practical viewpoint, for short constraint lengths also, convolutional codes considerably outperform block codes of the same order of complexity.

XI. Path Memory Truncation Metric Quantization and Synchronization

A major problem which arises in the implementation of a maximum likelihood decoder is the length of the path history which must be stored. In our previous discussion we ignored this important point and therefore implicitly assumed that all past data would be stored. A final decision was made by forcing the coder into a known (all zeros) state. We now remove this impractical condition. Suppose we truncate the path memories after M bits (branches) have been accumulated, by comparing all 2^K metrics for a maximum and deciding on the bit corresponding to that path (out of 2^K) with the highest metric M branches forward. If M is several times as large as K, the additional bit errors introduced in this way are very few, as we shall now demonstrate using the asymptotic results of the last section.

An additional bit error may occur due to memory truncation after M branches, if the bit selected is from an incorrect path which differed from the correct path M branches back and which has a higher metric, but which would ultimately be eliminated by the maximum likelihood decoder. But for a binary-tree code there can be no more than 2^M distinct paths which differ from the correct path M branches back. Of these we need concern ourselves only with those which have not merged with the correct path in the intervening nodes. As was originally shown by Forney [12], using the ensemble arguments of Section X we may bound the average probability of this event by [see (58)]

$$\tilde{P}_t < 2^{M\rho} 2^{-ME_0(\rho)/R}, \qquad 0 < \rho \leq 1. \tag{75}$$

To minimize this bound we should maximize the exponent $E_0(\rho)/R - \rho$ with respect to ρ on the unit interval. But this yields exactly $E_b(R)$, the upper bound exponent of (73) for block codes. Thus

$$\bar{P}_t < 2^{-ME_b(R)/R} \tag{76}$$

where $E_b(R)$ is the block coding exponent.

We conclude therefore that the memory truncation error is less than the bit error probability bound without truncation, provided the bound of (76) is less than the bound of (64). This will certainly be assured if

$$ME_b(R) > KE(R). \tag{77}$$

For very noisy channels we have from (69) and (74) or Fig. 17, that

$$\frac{M}{K} > \begin{cases} \dfrac{1}{1 - 2R/C}, & 0 \le R \le C/4 \\ \dfrac{1}{2(1 - \sqrt{R/C})^2}, & C/4 \le R \le C/2 \\ \dfrac{1 - R/C}{(1 - \sqrt{R/C})^2}, & C/2 < R < C. \end{cases}$$

For example, at $R = C/2$ this indicates that it suffices to take $M > (5.8)K$.

Another problem faced by a system designer is the amount of storage required by the metrics (or log-likelihood functions) for each of the 2^K paths. For a BSC this poses no difficulty since the metric is just the Hamming distance which is at most n, the number of code symbols, per branch. For the AWGN, on the other hand, the optimum metric is a real number, the analog output of a correlator, matched filter, or integrate-and-dump circuit. Since digital storage is generally required, it is necessary to quantize this analog metric. However, once the components y_{jk} of the optimum metric of (5), which are the correlator outputs, have been quantized to Q levels, the channel is no longer an AWGN channel. For biphase modulation, for example, it becomes a binary input Q-ary output discrete memoryless channel, whose transition probabilities are readily calculated as a function of the energy-to-noise density and the quantization levels. The optimum metric is not obtained by replacing y_{jk} by its quantized value $Q(y_{jk})$ in (5) but rather it is the log-likelihood function $\log P(\mathbf{y} \mid \mathbf{x}^{(m)})$ for the binary-input Q-ary-output channel.

Nevertheless, extensive simulation [24] indicates that for 8-level quantization even use of the suboptimal metric $\sum_k Q(y_{jk})x_{jk}^{(m)}$ results in a degradation of no more than 0.25 dB relative to the maximum likelihood decoder for the unquantized AWGN, and that use of the optimum metric is only negligibly superior to this. However, this is not the case for sequential decoding, where the difference in performance between optimal and suboptimal metrics is significant [11].

In a practical system other considerations than error performance for a given degree of decoder complexity often dictate the selection of a coding system. Chief among these are often the synchronization requirements. Convolutional codes utilizing maximum likelihood decoding are particularly advantageous in that no block synchronization is ever required. For block codes, decoding cannot begin until the initial point of each block has been located. Practical systems often require more complexity in the synchronization system than in the decoder. On the other hand, as we have by now amply illustrated, a maximum likelihood decoder for a convolutional code does not require any block synchronization because the coder is free running (i.e., it performs identical operations for each successive input bit and does not require that K bits be input before generating an output). Furthermore, the decoder does not require knowledge of past inputs to start decoding; it may as well assume that all previous bits were zeros. This is not to say that initially the decoder will operate as well, in the sense of error performance, as if the preceding bits of the correct path were known. On the other hand, consider a decoder which starts with an initially known path but makes an error at some point and excludes the correct path. Immediately thereafter it will be operating as if it had just been turned on with an unknown and incorrectly chosen previous path history. That this decoder will recover and stop making errors within a finite number of branches follows from our previous discussions in which it was shown that, other than for catastrophic codes, error sequences are always finite. Hence our initially unsynchronized decoder will operate just like a decoder which has just made an error and will thus always achieve synchronization and generally will produce correct decisions after a limited number of initial errors. Simulations have demonstrated that synchronization generally takes no more than four or five constraint lengths of received symbols.

Although, as we have just shown, branch synchronization is not required, code symbol synchronization within a branch is necessary. Thus, for example, for a binary-tree rate $R = 1/2$ code, we must resolve the two-way ambiguity as to where each two code-symbol branch begins. This is called node synchronization. Clearly if we make the wrong decisions, errors will constantly be made thereafter. However, this situation can easily be detected because the mismatch will cause *all* the path metrics to be small, since in fact there will not be any correct path in this case. We can thus detect this event and change our decision as to node synchronization (cf. Heller and Jacobs [24]). Of course, for an $R = 1/n$ code, we may have to repeat our choice n times, once for each of the symbols on a branch, but since n represents the redundancy factor or bandwidth expansion, practical systems rarely use $n > 4$.

XII. Other Decoding Algorithms for Convolutional Codes

This paper has treated primarily maximum likelihood decoding of convolutional codes. The reason for this was two-fold: 1) maximum likelihood decoding is closely related to the structure of convolutional codes and its consideration enhances our understanding of the ultimate capabilities, performance, and limitation of these codes; 2) for reasonably short constraint lengths ($K < 10$) its implementation is quite feasible[18] and worthwhile because of its optimality. Furthermore for $K \leq 6$, the complexity of maximum likelihood decoding is sufficiently limited that a completely parallel implementation (separate metric calculators) is possible. This minimizes the decoding time per bit and affords the possibility of extremely high decoding speeds [24].

Longer constraint lengths are required for extremely low error probabilities at high rates. Since the storage and computational complexity are proportional to 2^K, maximum likelihood decoders become impractical for $K > 10$. At this point *sequential decoding* [2], [4], [6] becomes attractive. This is an algorithm which sequentially searches the code tree in an attempt to find a path whose metric rises faster than some predetermined, but variable, threshold. Since the difference between the correct path metric and any incorrect path metric increases with constraint length, for large K generally the correct path will be found by this algorithm. The main drawback is that the number of incorrect path branches, and consequently the computation complexity, is a random variable depending on the channel noise. For $R < R_0$, it is shown that the average number of incorrect branches searched per decoded bit is bounded [6], while for $R > R_0$ it is not; hence R_0 is called the computational cutoff rate. To make storage requirements reasonable, it is necessary to make the decoding speed (branches/s) somewhat larger than the bit rate, thus somewhat limiting the maximum bit rate capability. Also, even though the average number of branches searched per bit is finite, it may sometimes become very large, resulting in a storage overflow and consequently relatively long sequences being erased. The stack sequential decoding algorithm [7], [18] provides a very simple and elegant presentation of the key concepts in sequential decoding, although the Fano algorithm [4] is generally preferable practically.

For a number of reasons, including buffer size requirements, computation speed, and metric sensitivity, sequential decoding of data transmitted at rates above about 100 K bits/s is practical only for hard-quantized binary received data (that is, for channels in which a hard decision -0 or $1-$ is made for each demodulated symbol). For the biphase modulated AWGN channel, of course, hard quantization (2 levels or 1 bit) results in an efficiency loss of approximately 2 dB compared with soft quantization (8 or more levels—3 or more bits). On the other hand, with maximum likelihood decoding, by employing a parallel implementation, short constraint length codes ($K \leq 6$) can be decoded at very high data rates (10 to 100 Mbits/s) even with soft quantization. In addition, the insensitivity to metric accuracy and simplicity of synchronization render maximum likelihood decoding generally preferable when moderate error probabilities are sufficient. In particular, since sequential decoding is limited by the overflow problem to operate at code rates somewhat below R_0, it appears that for the AWGN the crossover point above which maximum likelihood decoding is preferable to sequential decoding occurs at values of P_B somewhere between 10^{-3} and 10^{-5}, depending on the transmitted data rate. As the data rate increases the P_B crossover point decreases.

A third technique for decoding convolutional codes is known as *feedback decoding,* with threshold decoding [3] as a subclass. A feedback decoder basically makes a decision on a particular bit or branch in the decoding tree or trellis based on the received symbols for a limited number of branches beyond this point. Even though the decision is irrevocable, for limited constraint lengths (which are appropriate considering the limited number of branches involved in a decision) errors will propagate only for moderate lengths. When transmission is over a binary symmetric channel, by employing only codes with certain algebraic (orthogonal) properties, the decision on a given branch can be based on a linear function of the received symbols, called the *syndrome,* whose dimensionality is equal to the number of branches involved in the decision. One particularly simple decision criterion based on this syndrome, referred to as *threshold decoding,* is mechanizable in a very inexpensive manner. However, feedback decoders in general, and threshold decoders in particular, have an error-correcting capability equivalent to very short constraint length codes and consequently do not compare favorably with the performance of maximum likelihood or sequential decoding.

However, feedback decoders are particularly well suited to correcting error bursts which may occur in fading channels. Burst errors are generally best handled by using interleaved codes: that is, employing L convolutional codes so that the jth, $(L + j)$th $(2L + j)$th, etc., bits are encoded into one code for each $j = 0, 1, \cdots, L - 1$. This will cause any burst of length less than L to be broken up into random errors for the L independently operating decoders. Interleaving can be achieved by simply inserting $L - 1$ stage delay lines between stages of the convolutional encoder; the resulting single encoder then generates the L interleaved codes. The significant advantage of a feedback or threshold decoder is that the same technique can be employed in the decoder resulting in a single (time-shared) decoder rather than L decoders, providing feasible implementations for hard-quantized channels, even for protection against error bursts of thousands of bits. Details of feedback decoding

[18] Performing metric calculations and comparisons serially.

are treated extensively in Massey [3], Gallager [14], and Lucky *et al.* [16].

Appendix I

Generating Function for Structure of a Binary-Tree Convolutional Code for Arbitrary K and Error Bounds for Orthogonal Codes

We derive here the distance-invariant ($D = 1$) generating function $T(L, N)$ for any binary tree ($b = 1$) convolutional code of arbitrary constraint length K. It is most convenient in the general case to begin with the finite-state machine state-transition matrix for the linear equations among the state (node) variables. We exhibit this in terms of N and L for a $K = 4$ code as follows:

$$\begin{bmatrix} 1 & 0 & 0 & -NL & 0 & 0 & 0 \\ -L & 1 & 0 & 0 & -L & 0 & 0 \\ -NL & 0 & 1 & 0 & -NL & 0 & 0 \\ 0 & -L & 0 & 1 & 0 & -L & 0 \\ 0 & -NL & 0 & 0 & 1 & -NL & 0 \\ 0 & 0 & -L & 0 & 0 & 1 & -L \\ 0 & 0 & -NL & 0 & 0 & 0 & 1-NL \end{bmatrix} \cdot \begin{bmatrix} X_{001} \\ X_{010} \\ X_{011} \\ X_{100} \\ X_{101} \\ X_{110} \\ X_{111} \end{bmatrix} = \begin{bmatrix} NL \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \cdot \tag{78}$$

This pattern can be easily seen to generalize to a $2^{K-1} - 1$ dimensional square matrix of this form for any binary-tree code of constraint length K, and in general the generating function

$$T(L, N) = LX_{100\cdots0},$$

where $100 \cdots 0$ contains $(K - 2)$ zeros. (79)

From this general pattern it is easily shown that the matrix can be reduced to a dimension of 2^{K-2}. First combining adjacent rows, from the second to the last, pairwise, one obtains the set of $2^{K-2} - 1$ relations

$$NX_{i_1 i_2 \cdots i_{K-2} 0} = X_{i_1 i_2 \cdots i_{K-2} 1} \tag{80}$$

where $j_1, j_2, \cdots, j_{K-2}$ runs over all binary vectors except for the all zeros. Substitution of (80) into (78) yields a 2^{K-2}-dimensional matrix equation. The result for $K = 4$ is

$$\begin{bmatrix} 1 & 0 & -L & 0 \\ -NL & 1 & -NL & 0 \\ 0 & -L & 1 & -L \\ 0 & -NL & 0 & 1-NL \end{bmatrix} \cdot \begin{bmatrix} X_{001} \\ X_{011} \\ X_{101} \\ X_{111} \end{bmatrix} = \begin{bmatrix} NL \\ 0 \\ 0 \\ 0 \end{bmatrix} \cdot \tag{81}$$

Defining the new variable

$$X'_{00\cdots01} = NL\,X_{00\cdots01} + X_{00\cdots11} \tag{82}$$

(which corresponds to adding the second row to NL times the first), we obtain finally a $2^{K-2} - 1$ dimensional matrix equation, which for $K = 4$ is

$$\begin{bmatrix} 1 & -N(L+L^2) & 0 \\ -L & 1 & -L \\ -NL & 0 & 1-NL \end{bmatrix} \cdot \begin{bmatrix} X'_{001} \\ X_{101} \\ X_{111} \end{bmatrix} = \begin{bmatrix} N^2L^2 \\ 0 \\ 0 \end{bmatrix} \cdot \tag{83}$$

Note that (83) is the same as (78) for K reduced by unity, but with modifications in two places, both in the first row; namely, the first component on the right side is squared, and the middle term of the first row is reduced by an amount NL^2. Although we have given the explicit result only for $K = 4$, it is easily seen to be valid for any K.

Since in all respects, except these two, the matrix after this sequence of reductions is the same as the original but with its dimension reduced corresponding to a reduction of K by unity, we may proceed to perform this sequence of reductions again. The steps will be the same except that now in place of (80), we have

$$NX_{i_1 i_2 \cdots i_{K-3} 01} = X_{i_1 i_2 \cdots i_{K-3} 11} \tag{80'}$$

and in place of (82)

$$X''_{00\cdots01} = NL\,X'_{00\cdots01} + X_{00\cdots111} \tag{82'}$$

while in place of (81) the right of center term of the first row is $-(L + L^2)$ and the first component on the right side is N^2L^2. Similarly in place of (83) the center term of the first row is $-N(L + L^2 + L^3)$ and the first component on the right side is N^3L^3.

Performing this sequence of reductions $K - 2$ times in all, but omitting the last step—leading from (81) to (83)—in the last reduction, the original $2^{K-1} - 1$ equations are reduced in the general case to the two equations

$$\begin{bmatrix} 1 & -(L + L^2 + \cdots L^{K-2}) \\ -NL & 1 - NL \end{bmatrix} \cdot \begin{bmatrix} X_{00-01}{}^{(K-3)} \\ X_{11\cdots1} \end{bmatrix} = \begin{bmatrix} (NL)^{K-2} \\ 0 \end{bmatrix} \tag{84}$$

whence it follows that

$$X_{11\cdots1} = \frac{(NL)^{K-1}}{1 - N(L + L^2 + \cdots + L^{K-1})} \tag{85}$$

Applying (79) and the $K - 2$ extensions of (80) and (80′) we find

$$\begin{aligned} T(L, N) &= LX_{100\cdots00} = LN^{-1}X_{100\cdots01} \\ &= LN^{-2}X_{100\cdots011} = \cdots = LN^{-(K-2)}X_{11\cdots1} \\ &= \frac{NL^K}{1 - N(L + L^2 + \cdots + L^{K-1})} \\ &= \frac{NL^K(1 - L)}{1 - L(1 + N) + NL^K} \end{aligned} \tag{86}$$

If we require only the path length structure, and not the number of bit errors corresponding to any incorrect path, we may set $N = 1$ in (86) and obtain

$$T(L) = \frac{L^K}{1 - (L + L^2 + \cdots + L^{K-1}} = \frac{L^K(1 - L)}{1 - 2L + L^K}. \tag{87}$$

If we denote as an upper bound an expression which is the generating function of more paths than exist in our state diagram, we have

$$T(L) < \frac{L^K}{1 - 2L}. \tag{88}$$

As an additional application of this generating function technique, we now obtain bounds on P_E and P_B for the class of orthogonal convolutional (tree) codes introduced by Viterbi [10]. For this class of codes, to each of the 2^K branches of the K-state diagram there corresponds one of 2^K orthogonal signals. Given that each signal is orthogonal to all others in $n \geq 1$ dimensions, corresponding to n channel symbols or transmission times (as, for example, if each signal consists of n different pulses out of $2^K n$ possible positions), then the weight of each branch is n. Consequently, if we replace L, the path length enumerator, by D^n in (86) we obtain for orthogonal codes

$$T(D, N) = \frac{ND^{nK}(1 - D^n)}{1 - D^n(1 + N) + ND^{nK}}. \tag{89}$$

Then using (48) and (49), the first-event error probability for orthogonal codes is bounded by

$$P_E < \frac{{D_0}^{nK}(1 - {D_0}^n)}{1 - 2{D_0}^n + {D_0}^{nK}} < \frac{{D_0}^{nK}(1 - {D_0}^n)}{1 - 2{D_0}^n} \tag{90}$$

and the bit error probability bound is

$$\begin{aligned} P_B &< \left.\frac{dT(N, D)}{dN}\right|_{N=1, D=D_0} \\ &= \frac{{D_0}^{nK}(1 - {D_0}^n)^2}{(1 - 2{D_0}^n + {D_0}^{nK})^2} < \frac{{D_0}^{nK}(1 - {D_0}^n)^2}{(1 - 2{D_0}^n)^2} \end{aligned} \tag{91}$$

where D_0 is a function of the channel transition probabilities or energy-to-noise ratio and is given by (46).

Acknowledgment

The author gratefully acknowledges the considerable stimulation he has received over the course of writing the several versions of this paper from Dr. J. A. Heller, whose recent work strongly complements and enhances this effort, for numerous discussions and suggestions and for his assistance in its presentation at the Linkabit Corporation "Seminars on Convolutional Codes." This tutorial approach owes part of its origin to Dr. G. D. Forney, Jr., whose imaginative and perceptive reinterpretation of my original work has aided immeasurably in rendering it more comprehensible. Also, thanks are due to Dr. J. K. Omura for his careful and detailed reading and correction of the manuscript during his presentation of this material in the UCLA graduate course on information theory.

References

[1] P. Elias, "Coding for noisy channels," in *1955 IRE Nat. Conv. Rec.*, vol. 3, pt. 4, pp. 37–46.

[2] J. M. Wozencraft, "Sequential decoding for reliable communication," in *1957 IRE Nat. Conv. Record*, vol. 5, pt. 2, pp. 11–25.

[3] J. L. Massey, *Threshold Decoding*. Cambridge, Mass.: M.I.T. Press, 1963.

[4] R. M. Fano, "A heuristic discussion of probabilistic decoding," *IEEE Trans. Inform. Theory*, vol. IT-9, Apr. 1963, pp. 64–74.

[5] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inform. Theory*, vol. IT-11, Jan. 1965, pp. 3–18.

[6] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.

[7] K. S. Zigangirov, "Some sequential decoding procedures," *Probl. Peredach Inform.*, vol. 2, no. 4, 1966, pp. 13–25.

[8] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels," *Inform. Contr.*, vol. 10, 1967, pt. I, pp. 65–103, pt. II, pp. 522–552.

[9] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, Apr. 1967, pp. 260–269.

[10] ——, "Orthogonal tree codes for communication in the presence of white Gaussian noise," *IEEE Trans. Commun. Technol.*, vol. COM-15, April 1967, pp. 238–242.

[11] I. M. Jacobs, "Sequential decoding for efficient communication from deep space," *IEEE Trans. Commun. Technol.*, vol. COM-15, Aug. 1968, pp. 492–501.

[12] G. D. Forney, Jr., "Coding system design for advanced solar missions," submitted to NASA Ames Res. Ctr. by Codex Corp., Watertown, Mass., Final Rep., Contract NAS2-3637, Dec. 1967.

[13] J. L. Massey and M. K. Sain, "Inverses of linear sequential circuits," *IEEE Trans. Comput.*, vol. C-17, Apr. 1968, pp. 330–337.

[14] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[15] T. N. Morrissey, "Analysis of decoders for convolutional codes by stochastic sequential machine methods," Univ. Notre Dame, Notre Dame, Ind., Tech. Rep. EE-682, May 1968.

[16] R. W. Lucky, J. Salz, and E. J. Weldon, *Principles of Data Communication*. New York: McGraw-Hill, 1968.

[17] J. K. Omura, "On the Viterbi decoding algorithm," *IEEE Trans. Inform. Theory,* vol. IT-15, Jan. 1969, pp. 177–179.
[18] F. Jelinek, "Fast sequential decoding algorithm using a stack," *IBM J. Res. Dev.,* vol. 13, no. 6, Nov. 1969, pp. 675–685.
[19] E. A. Bucher and J. A. Heller, "Error probability bounds for systematic convolutional codes," *IEEE Trans. Inform. Theory,* vol. IT-16, Mar. 1970, pp. 219–224.
[20] J. P. Odenwalder, "Optimal decoding of convolutional codes," Ph.D. dissertation, Dep. Syst. Sci., Sch. Eng. Appl. Sci., Univ. California, Los Angeles, 1970.
[21] G. D. Forney, Jr., "Coding and its application in space communications," *IEEE Spectrum,* vol. 7, June 1970, pp. 47–58.
[22] ——, "Convolutional codes I: Algebraic structure," *IEEE Trans. Inform. Theory,* vol. IT-16, Nov. 1970, pp. 720–738; "II: Maximum likelihood decoding," and "III: Sequential decoding," *IEEE Trans. Inform. Theory,* to be published.
[23] W. J. Rosenberg, "Structural properties of convolutional codes," Ph.D. dissertation, Dep. Syst. Sci., Sch. Eng. Appl. Sci., Univ. California, Los Angeles, 1971.
[24] J. A. Heller and I. M. Jacobs, "Viterbi decoding for satellite and space communication," this issue, pp. 835–848.
[25] A. R. Cohen, J. A. Heller, and A. J. Viterbi, "A new coding technique for asynchronous multiple access communication," this issue, pp. 849–855.

Andrew J. Viterbi (S'54–M'58–SM'63) was born in Bergamo, Italy, on March 9, 1935. He received the B.S. and M.S. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1957, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1962.

While attending M.I.T., he participated in the cooperative program at the Raytheon Company. In 1957 he joined the Jet Propulsion Laboratory where he became a Research Group Supervisor in the Communications Systems Research Section. In 1963 he joined the faculty of the University of California, Los Angeles, as an Assistant Professor. In 1965 he was promoted to Associate Professor and in 1969 to Professor of Engineering and Applied Science. He was a cofounder in 1968 of Linkabit Corporation of which he is presently Vice President.

Dr. Viterbi is a member of the Editorial Boards of the PROCEEDINGS OF THE IEEE and of the journal *Information and Control.* He is a member of Sigma Xi, Tau Beta Pi, and Eta Kappa Nu and has served on several governmental advisory committees and panels. He is the coauthor of a book on digital communication and author of another on coherent communication, and he has received three awards for his journal publications.

Raffaello, La Scuola d'Atene, particolare con Euclide e gli studenti di geometria.
Stanza della Segnatura (Città del Vaticano).

Le telecomunicazioni verso l'assetto numerico

Saluto dell'AIIT

Giovanni De Guzzis: Signore e Signori, come presidente dell'AIIT ho l'onore di porgervi il benvenuto a questo seminario sulle prospettive delle telecomunicazioni che si svolge immediatamente prima del conferimento della laurea Honoris Causa al Professor Andrew J. Viterbi da parte dell'Università di Tor Vergata.

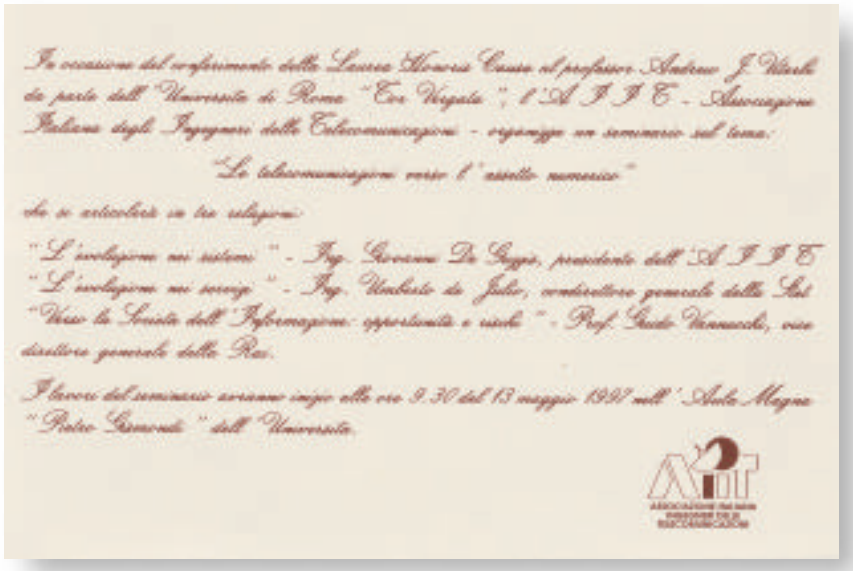
L'AIIT, Associazione Italiana Ingegneri delle Telecomunicazioni, ha trentacinque anni di vita e riunisce la maggior parte di coloro che operano nel settore delle telecomunicazioni, settore che in Italia ha avuto un enorme sviluppo: oggi infatti tutte le famiglie sono collegate alla rete telefonica e un adulto su sette dispone di un telefono cellulare.

Esistono in Italia organizzazioni analoghe; l'AIIT credo però che abbia caratteristiche che la distinguono dalle altre: cerca infatti di dare molto rilievo agli aspetti legati ai contatti umani e sociali, perché riunisce persone più che enti o istituzioni. L'Associazione ha anche tra gli obiettivi fondamentali la promozione di riunioni come quella alla quale oggi prendiamo parte, che perseguono lo scopo di diffondere l'informazione culturale e tecnica.

Altri con maggior profondità e conoscenza della mia illustreranno la vita e l'attività del Professor Viterbi. Io mi limito a segnalare che pochi come lui hanno contribuito a disegnare il mondo in cui viviamo: la conquista dello spazio, la televisione numerica, le comunicazioni digitali sono tutti campi nei quali questo illustre scienziato ha dato un contributo determinante; eppure il suo nome non è forse abbastanza conosciuto da milioni di italiani, che ogni giorno traggono vantaggio dalle sue intuizioni e dalle sue scoperte. Spero quindi che questa sia l'occasione perché gli italiani conoscano meglio un personaggio di così grande rilievo.

L'Associazione ha ritenuto opportuno approfittare di questa occasione per organizzare al tempo stesso una giornata di studio su "Le telecomunicazioni verso l'assetto numerico"; e ho accettato con piacere di tenere la prima relazione riguardante l'"evoluzione dei sistemi".

È un argomento di grande complessità sul quale vorrei quindi dirvi il mio punto di vista, tenendo presente che in scenari in così rapida evoluzione le idee e gli spunti possono servire prima per una riflessione generale e poi per un maggiore approfondimento.



Le telecomunicazioni verso l'assetto numerico

L'evoluzione nei sistemi

GIOVANNI DE GUZZIS

Giovanni De Guzzis, Amministratore Delegato e Direttore Generale della Ericsson, durante il suo intervento.

La tecnologia

Qual è stata la spinta determinante che ha portato la numerizzazione nel campo delle telecomunicazioni?

Non è facile dare una risposta a questa domanda perché l'impatto della numerizzazione in questo settore è stato così profondo e rivoluzionario che diversi aspetti e conseguenze possono essere enfatizzati a seconda della visuale da cui ci si pone: affidabilità della gestione, intelligenza del controllo, qualità della trasmissione, nuovi servizi, integrazione e omogeneizzazione delle tecniche.

Forse anche la mia è un'opinione che deriva da una visuale parziale; ma sono profondamente convinto che la numerizzazione della rete di telecomunicazione ha consentito come risultato fondamentale l'utilizzo efficiente e ottimale delle risorse tecnologiche di base e soprattutto di quelle trasmissive, e ha reso disponibile agli utenti la capacità di scambiare informazioni in quantità sempre crescente. Questa opinione è particolarmente evidente nel caso della trasmissione radio dove, grazie alla numerizzazione, la capacità C della rete mobile[1] è potuta crescere continuamente a parità dell'esiguo spettro radio disponibile.

Ma i risultati ottenuti per qualsiasi altro portante fisico sono stati parimenti straordinari e il limite di Shannon ($C=B \log_2 (1+S/N)$), una volta considerato irraggiungibile, viene oggi di fatto sfiorato in parecchie applicazioni.

La tecnologia del trattamento dell'informazione numerica ha fatto in questi ultimi anni passi da gigante. La densità dell'hardware è aumentata in misura esponenziale: le dimensioni del transistor hanno raggiunto lo 0,1 μm e il chip di memoria DRAM da 1 Gbit (capace cioè di memorizzare un miliardo di bit) è ormai alle porte.

La tecnologia CMOS da 0,25 μm si sta producendo nelle fonderie di silicio più avanzate.

La velocità di elaborazione è cresciuta di conseguenza in modo vertiginoso: è stato annunciato un nuovo processore con un orologio superiore a 500 MHz che permetterà in tempo reale la codifica video in MPEG2.

Il costo del bit elaborato è quindi continuamente diminuito e si ridurrà in futuro in modo ancora più accentuato (si veda la figura 1 e la tabella 1).

Gli algoritmi di codifica di canale, anche i più complessi, possono essere utilizzati a basso costo.

(1) *C=Erlang/(km² x MHz).*

Data	*Sistema*	*MIPS*	*Costo in $USA (x 1000)*	*Costo in $USA per MIPS*
1975	IBM Mainframe	10	10000	1000000
1976	Cray 1	160	20000	125000
1979	Digital VAX	1	200	200000
1981	IBM PC	0,25	3	12000
1984	Sun 2	1	10	10000
1994	Pentium Chip PC	66	3	45,5
1996	Sony PCX Video Game	500	0,5	1
1997	Microunity Set Top Box	1000	0,5	0,5

MIPS = Milioni di Istruzioni Per Secondo

Tabella 1 *Evoluzione dei sistemi e dei costi.*

Allo stesso tempo gli algoritmi di riduzione di ridondanza hanno ridotto la necessità di banda per segnali sia vocali che video.

Il segnale video PAL (campionato allo standard di 13,5 MHz) richiederebbe una frequenza di cifra di 166 Mbit/s; ma applicando l'algoritmo di compressione MPEG2 (broadcast quality) sono necessari solo 6 Mbit/s. Il che significa che sugli 8 MHz di banda base del segnale video PAL analogico (con codifica a 64 QAM) possono essere trasmessi sei canali televisivi equivalenti.

Vale la pena di ricordare che se ci si accontenta di una qualità più ridotta, quale quella del segnale VHS, con la compressione MPEG1 sono necessari solo 1,2 Mbit/s (e quindi sul canale PAL sopra citato sarebbero contenuti trenta canali televisivi con "la qualità di un videoregistratore").

Inoltre sulla lunga distanza il costo della trasmissione per bit è ridotto praticamente a zero dall'introduzione delle fibre ottiche e dalle tecniche di modulazione WDM (è stata sperimentata la trasmissione alla capacità di un Tbit/s su una fibra ottica convenzionale[2] e su una distanza di 150 km modulando 55 portanti

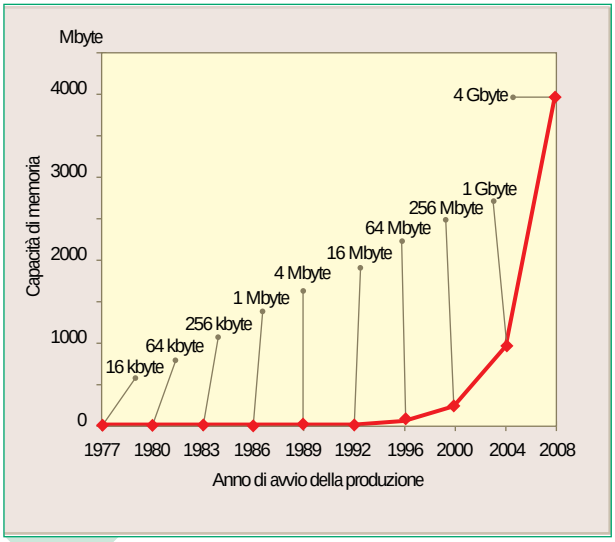


Figura 1 *Evoluzione della capacità di memoria dei semiconduttori.*

ottiche a 20 Gbit/s).

Naturalmente lo sviluppo della tecnologia non ha avuto solo effetti sulla trasmissione ma ha permesso di introdurre tecniche di commutazione (asincrone) più sofisticate ed efficienti, orientate alla totale ed efficace condivisione delle risorse di rete a costi sempre più competitivi (un esempio è quello riguardante l'impiego delle tecnologie ATM).

Il fenomeno Internet

Grazie allo sviluppo descritto per le tecniche e per le tecnologie, l'utente residenziale della rete telefonica tradizionale, con una spesa di circa 100 dollari, ha potuto acquistare per il suo computer dei modem disponibili nel mercato commodity (cioè in negozi specializzati) che gli consentono di effettuare collegamenti dati sulla

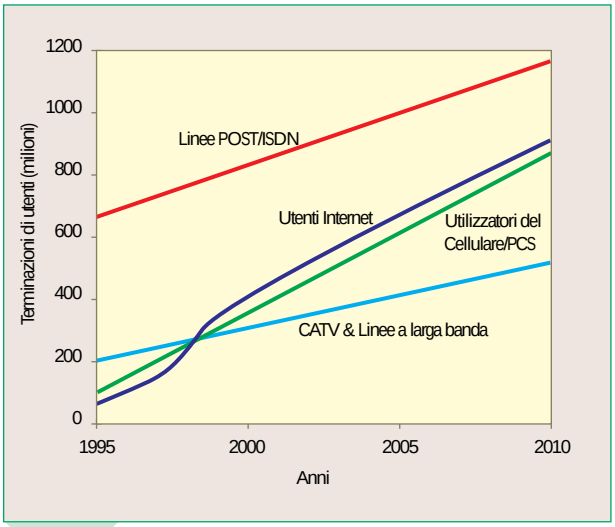


Figura 2 *Previsione attuale di crescita per le differenti reti e servizi.*

normale linea commutata a 28,8 kbit/s, a 36,6 kbit/s e più di recente a 56,5 kbit/s.

La disponibilità per l'utente di questa capacità trasmissiva a basso costo ha scatenato il fenomeno Internet; è ormai grande l'aspettativa che si sta creando nelle famiglie di poter disporre di servizi di comunicazione telematica, di information retrieval[3], di posta elettronica e di spettacolo.

Il tasso di crescita dell'utenza Internet è superiore a quello del radiomobile (figura 2): sono previsti

(2) *Cento miliardi di bit/s, Terabit/s, Tbit/s.*

(3) *Estrazione di informazioni.*

500 milioni di utenti Internet nel Duemila. I fornitori di servizi per la rete Internet stanno dando una risposta pronta a queste richieste manifestate dall'utenza.

I siti WEB della rete sono passati da 23.500 a 230.000 da giugno '95 a giugno '96.

Se si tiene conto che la mobilità è l'altra fondamentale esigenza che il cliente residenziale richiede che venga soddisfatta (sono previsti 400 milioni di utenti nel Duemila) il binomio domanda di mobilità-accesso a Internet sembra costituire la sfida che i sistemi di comunicazione del futuro dovranno fronteggiare.

La rete di interconnessione

Sulla base di quanto si è esposto finora si pone il quesito: come evolveranno le reti di telecomunicazione per soddisfare queste esigenze via via crescenti?

Di fatto esistono due tipi di reti: la prima telefonica (fissa o mobile), l'altra per dati.

Queste due reti si trovano oggi, per quanto si è detto prima, a condividere l'accesso dell'utenza residenziale. Più precisamente, la rete telefonica, per un fenomeno che i gestori non hanno guidato e su cui finora non hanno investito, si è trovata a dover fornire ai propri clienti l'accesso alla rete dati (figura 3).

Le due reti si sono poi evolute in modo sostanzialmente autonomo utilizzando tecnologie completamente diverse e basandosi su architetture anch'esse differenti: l'architettura della rete telefonica è chiamata *Telephone Service Control Architecture*; l'altra è denominata *Computer Network Architecture*.

La prima è ad intelligenza interna sempre più centralizzata sia per la fornitura di servizi (ad esempio la rete intelligente) sia per la gestione e per la manutenzione (OSS); essa permette l'interconnessione, fornisce servizi ed è caratterizzata da una elevata qualità, affidabilità e disponibilità.

La seconda è un'architettura completamente distribuita, basata su instradatori (router), fornisce essenzialmente il servizio di trasporto e demanda qualità e affidabilità alle unità periferiche e la disponibilità alla ridondanza. L'intelligenza non risiede affatto in rete ma si situa ai suoi bordi. L'estrema efficienza del protocollo IP, il basso costo e l'*HTTP (Hyper/Text Transport Protocol)* hanno di fatto decretato il successo di essa e hanno in tal modo favorito il fenomeno Internet.

I servizi di information retrieval e di posta elettronica sono oggi totale appannaggio di questa rete.

I servizi in voce e video in tempo reale non sono invece oggi gestiti adeguatamente da questo tipo di architettura e probabilmente questa limitazione si manterrà nel tempo indipendentemente dall'evoluzione degli specifici protocolli e dalla banda che sarà resa disponibile agli utenti.

Inoltre la mobilità si presta meglio a essere gestita da un'architettura a controllo e con intelligenza centralizzata, tipica dell'altra rete.

Cosa succederà in futuro? Sarà possibile una convergenza tecnologica e architetturale delle due reti? Cosa preferiranno i clienti?

La rete tradizionale, dedicata principalmente al trasporto del segnale fonico, dovrà competere con l'economicità della rete per la trasmissione di dati, mentre quella per i dati dovrà competere in affidabilità, qualità e disponibilità con quella tradizionale. Questa evoluzione porterà sicuramente a miglioramenti su entrambe le reti; ma, secondo la mia opinione, esse rimarranno sostanzialmente separate nella parte di interconnessione.

L'accesso alle due architetture di rete sarà invece comune e dovrà essere presente un intermediario che selezioni la rete di trasporto in funzione delle caratteristiche di economicità, qualità e affidabilità richieste dal servizio da erogare.

La semplificazione dell'interfaccia commerciale per l'utente può essere un veicolo di più facile diffusione dei servizi: un solo *service provider* che raggruppi tutta l'offerta messa a disposizione dell'utente.

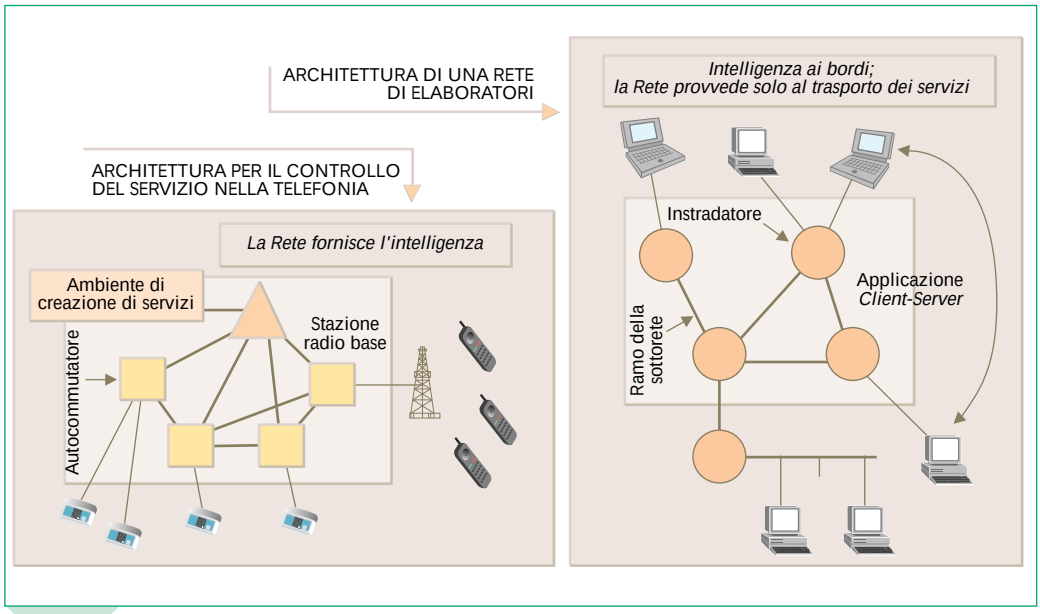


Figura 3 *Architettura delle due reti: telefonica e dati.*

L'accesso

Indipendentemente dalla struttura della rete di interconnessione, l'aumento della capacità trasmissiva all'accesso (sia mobile che fisso) per il singolo utilizzatore è il presupposto indispensabile per soddisfarne le esigenze presenti e stimolarne di nuove.

Più che la rete di interconnessione, sarà uno sviluppo adeguato della rete di accesso che consentirà lo sviluppo futuro delle telecomunicazioni.

dei doppini in Italia) oppure 2 Mbit/s e 0,5 Mbit/s per distanze fino a 5-6 km.

Questa soluzione non richiede grandi investimenti per l'infrastruttura di base, in quanto la capacità è fornita solo all'utente che ne faccia richiesta. L'investimento maggiore è quello relativo al singolo utilizzatore e con esso può essere pienamente soddisfatta l'esigenza di un utente medio Internet fino all'anno 2010.

Una soluzione che richiede più investimenti infrastrutturali è quella detta *FTTC (Fiber To The Curb)* in

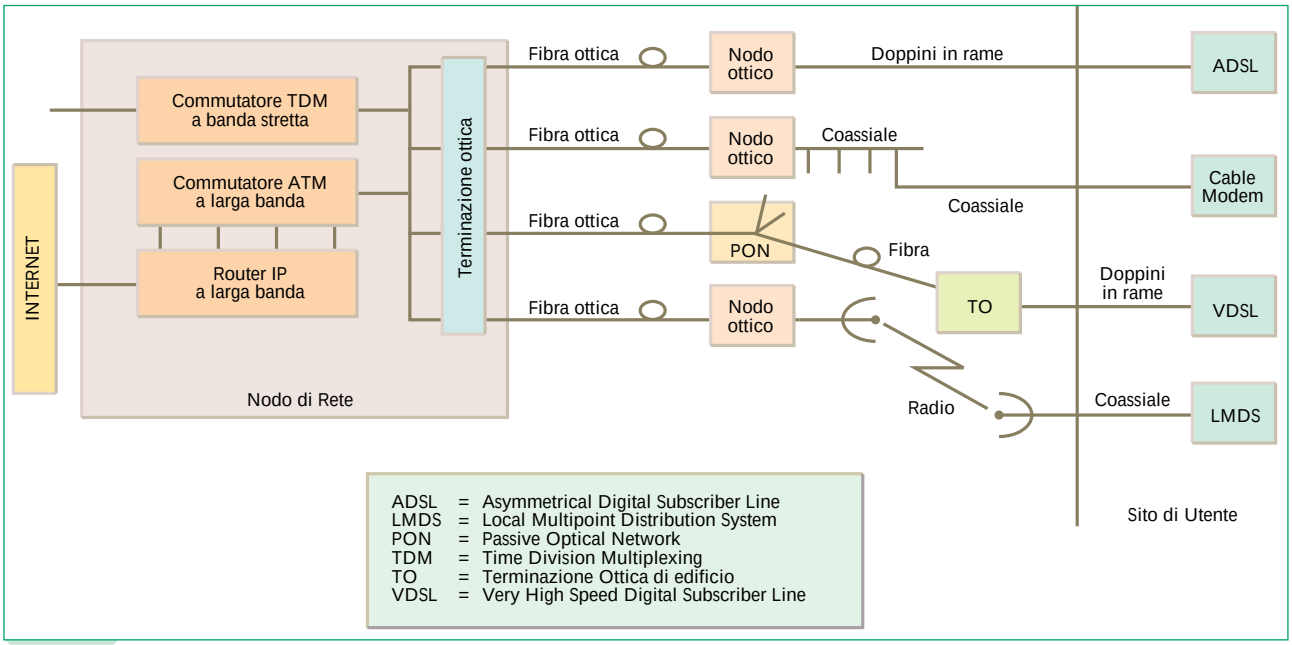


Figura 4 *Soluzioni possibili per l'accesso Internet a larga banda sull'ultimo miglio.*

Al singolo utente si potrà arrivare (figura 4):

- direttamente dalla centrale su doppino con tecniche *ADSL (Asymmetrical Digital Subscriber Line);*
- con fibra che sarà spinta quanto più possibile in prossimità della sede di utente, raggiungendo così un nodo ottico (*fiber node)* dal quale parte un coassiale;
- con fibra fino al fiber node dal quale parte un doppino in rame che impieghi la tecnica *VDSL (Very high speed Digital Subscriber Line);*
- con soluzioni radio *(wireless)* dal nodo ottico (o dalla centrale).

Le soluzioni radio sembrano essere le più promettenti perché permettono di condividere direttamente l'accesso tra più utenti e non richiedono costi di installazione di qualsiasi tipo di cavo.

Le altre soluzioni richiedono organi di concentrazione situati relativamente vicino alla centrale per la condivisione dell'accesso alla stessa centrale.

Per quanto riguarda l'accesso fisso si può incrementare la banda di accesso per l'utente utilizzando il doppino di rame già installato con tecniche ADSL che consentono di trasmettere segnali con banda fino a 6-8 Mbit/s nel senso centrale-terminazione di utente *(downstream)* e 1 Mbit/s nel senso inverso *(upstream)* per distanze fino a 3,5 km (che costituiscono il 95 per cento cui ci si avvicina all'utente raggiungendo con la fibra ottica un "fiber node" da cui o parte un cavo coassiale o un doppino che impieghi la tecnica VDSL, che su distanze di 300 m può permettere la trasmissione di flussi fino a 55 Mbit/s e su 100 m flussi a 155 Mbit/s.

L'accesso può anche essere realizzato con ponti radio in tecnica punto-multipunto *LMDS (Local Multipoint Distribution System)* con antenne paraboliche a 28-38 GHz, utilizzabile quando si ha visibilità tra le antenne.

Tutte queste soluzioni consentiranno non solo servizi di information retrieval, ma anche servizi video di vario genere.

L'utente di comunicazioni mobili non fruirà della stessa larghezza di banda, ma dovrà accontentarsi in futuro di 500-2000 kbit/s, che sono valori di velocità di cifra comunque ben maggiori di quelli permessi dai sistemi mobili oggi impiegati TACS (14,4 kbit/s), GSM (9,6 kbit/s) e DECT (32 kbit/s).

Oltre il Duemila: lo scenario fra tre possibilità estreme

Le tecnologie descritte possono essere di ausilio allo sviluppo di diversi scenari, ma l'evoluzione delle

telecomunicazioni sarà frutto soprattutto di alcune componenti diverse, quali (figura 5):
- ambiente economico e di regolamentazione;
- comportamento dei consumatori;
- sviluppo dei servizi.

Possiamo prevedere tre scenari limite: essi definiscono uno spazio di possibilità di evoluzione nell'ambito del quale molto probabilmente si collocherà il mondo delle telecomunicazioni dell'inizio del prossimo millennio; e sarà più vicino al primo, al secondo o al terzo scenario a seconda delle diverse influenze che sono già entrate in gioco ma che devono ancora chiarire gli effetti da essi prodotti.

In uno studio effettuato dalla Ericsson sono stati individuati tre scenari, chiamati rispettivamente:
- Gran Tradizione;
- Up and Away;
- Service Mania.

Il primo scenario: Gran Tradizione

In un contesto di sviluppo economico rallentato, con uno stile di consumi riflessivo e poco attento alle tecnologie innovative, con uno stile di politica economica rivolta al sostegno dell'occupazione e quindi favorevole a una introduzione lenta della concorrenza e delle privatizzazioni, si ipotizza l'affermarsi dello scenario denominato "Gran Tradizione" (figure 6 e 7).

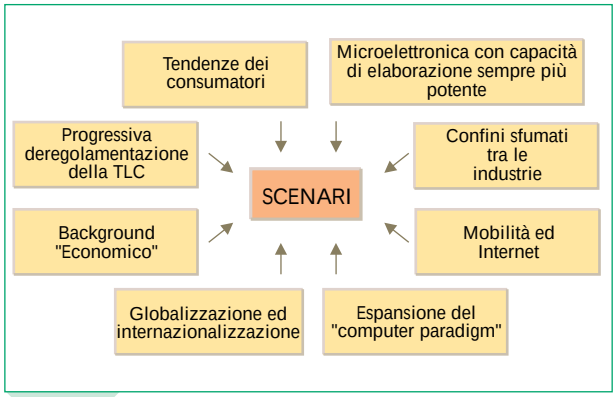


Figura 5 *Elementi guida per i diversi possibili scenari di evoluzione delle telecomunicazioni.*

Con esso è confermata l'architettura di rete più tradizionale, che vede dominante il paradigma telefonico. Essa è caratterizzata dalla massima centralizzazione dell'intelligenza di rete, gestita da grandi gestori di reti di telecomunicazioni, che sarebbero dunque gli animatori dell'innovazione e dell'introduzione dei nuovi servizi.

Mentre la rete mobile continua a espandersi sugli stessi ritmi degli anni Novanta, la rete fissa è dominata dalla commutazione tradizionale e da servizi centralizzati che generano una rilevante quota di entrate per i gestori.

La diffusione dei servizi ATM è lenta e così anche la nascita di servizi multimediali che rimangono al di sotto delle previsioni oggi più accreditate.

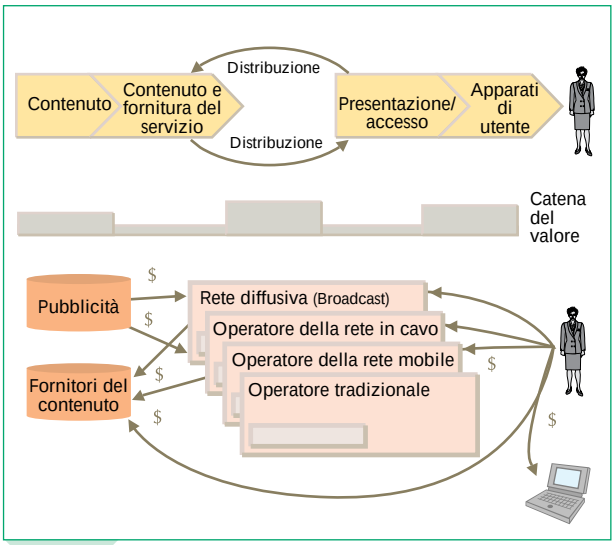


Figura 6 *Primo scenario: Gran Tradizione.*

I consumatori si evolvono poco sul piano della cultura tecnologica; la diffusione dei nuovi servizi è lenta ed è principalmente orientata dall'iniziativa dei gestori. Lo sviluppo delle telecomunicazioni avanzate rimane principalmente concentrato nei Paesi ad alto reddito (Stati Uniti di America, Europa e Giappone). Nel resto del mondo si diffondono solo servizi telefonici di base e per la mobilità.

Il secondo scenario: Up and Away

Nello scenario "Up and Away" (figure 8, 9 e 10) la

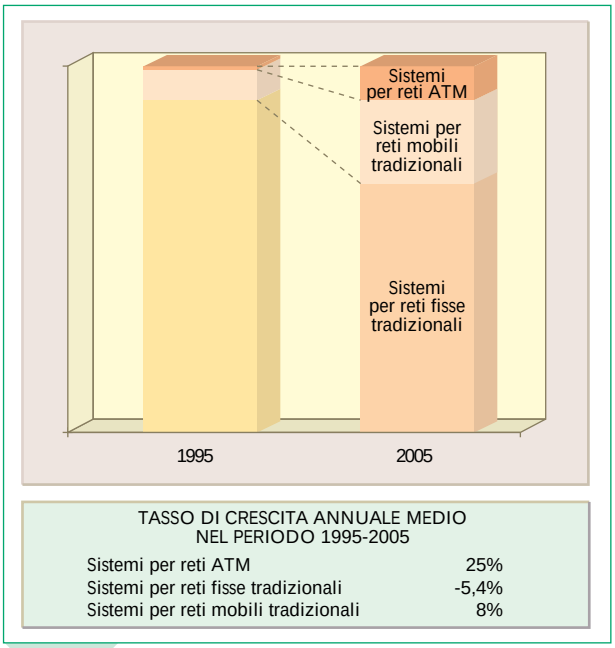


Figura 7 *Variazione del mercato della commutazione nello scenario Gran Tradizione.*

situazione che si presenta, è diametralmente opposta per quanto riguarda ruolo dei gestori tradizionali e maturità tecnologica dell'utenza.

Sulla base di un'elevata crescita economica e di un

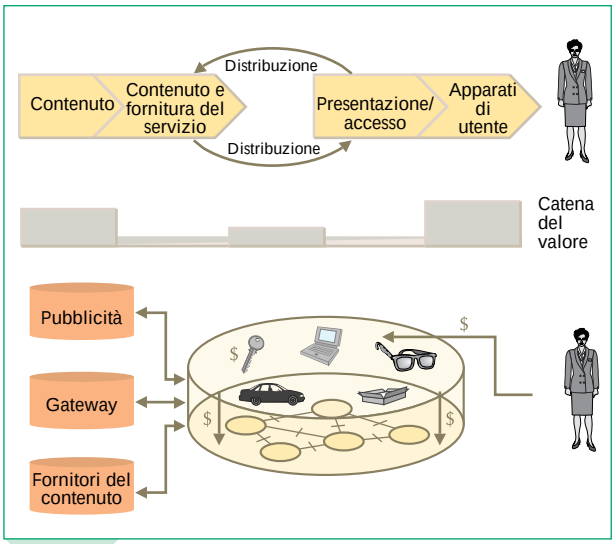


Figura 8 *Secondo scenario: Up and Away.*

elevato livello di consumi, la spinta all'accelerazione del consumo tecnologico è spontanea. L'ente di regolamentazione elimina dunque tutte le barriere che si oppongono all'affermazione della Società dell'Informazione.

La Computer Education è largamente diffusa nella società, tra i consumatori. La funzionalità dei

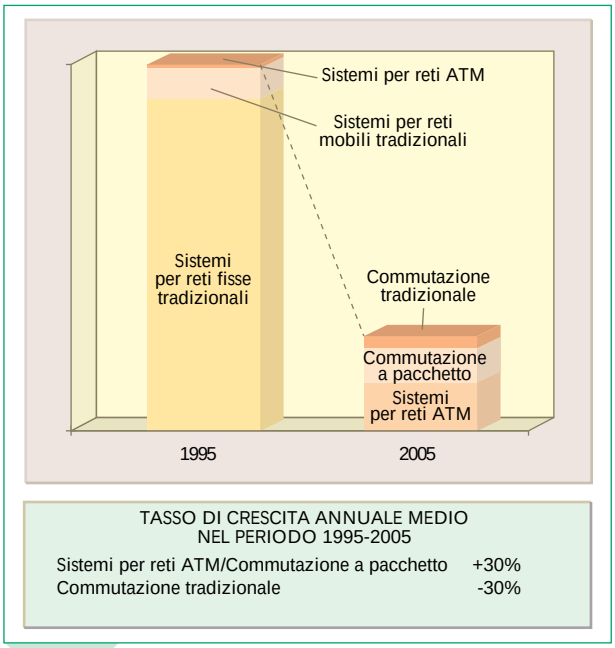


Figura 9 *Variazione negli anni del mercato della commutazione nello scenario Up and Away.*

servizi è controllata dalle unità periferiche (e residenziali) poste nei terminali degli utenti stessi. Domina il trasporto dell'informazione una rete, che possiamo chiamare Futurenet, e che rappresenta l'evoluzione di Internet. La tecnologia dominante è quella dei router. Gli standard prescritti sono deboli mentre l'affermazione di standard di fatto rappresenta la regola. Le tecnologie di compressione sono avanzatissime e la necessità di banda nei sistemi di accesso rimane stabile in virtù del progresso nell'efficienza d'impiego che se ne fa. Gli investimenti in infrastruttura sono minimi, mentre quelli per i sistemi di utente sono effettuati su misura per le necessità del singolo e sono quindi estremamente flessibili e differenziati. Lo sviluppo della Società dell'Informazione è affidato ai produttori di sistemi d'utente ed a società dedicate allo sviluppo Software oltreché alla maturità tecnologica e alla creatività degli utenti.

Il terzo scenario: Service Mania

Nel terzo scenario, "Service Mania", lo sviluppo economico è sostenuto, i consumi sono evoluti; tuttavia l'utenza, pur richiedendo servizi sofisticati, non

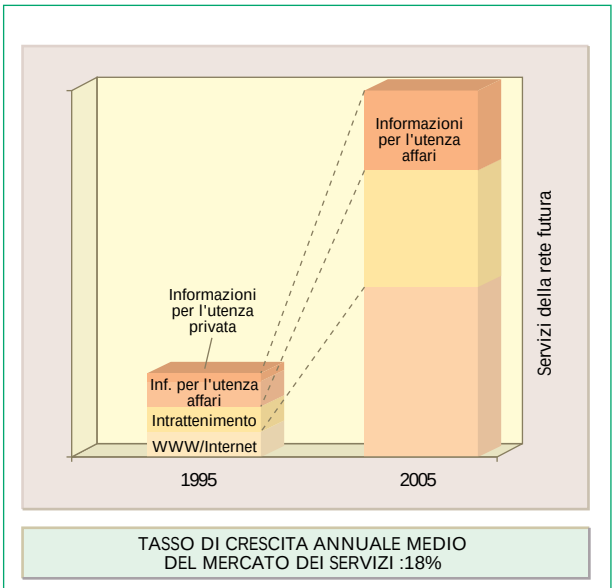


Figura 10 *Mercato dei servizi nello scenario Up and Away.*

raggiunge rapidamente un'elevata "Computer Education" (figure 11, 12 e 13).

L'ente di regolamentazione, per accelerare lo sviluppo della Società dell'Informazione, agevola l'ingresso di nuovi attori in concorrenza.

L'architettura di rete, che serve i bisogni di comunicazione della Società dell'Informazione, è più complessa e prevede la sopravvivenza di più reti che hanno il compito di fornire servizi specializzati. Mentre sopravvive la rete tradizionale per il servizio telefonico, si diffondono allo stesso tempo reti dati a

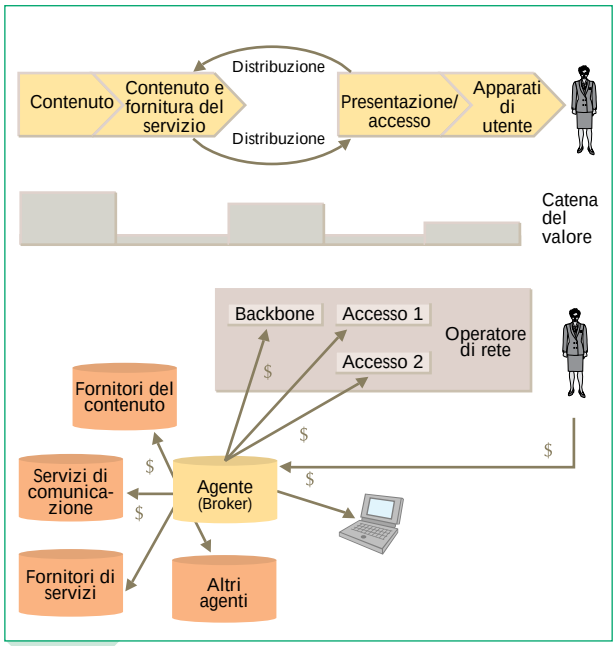


Figura 11 *Terzo scenario: Service Mania.*

larga banda che impiegano la tecnologia ATM.

Agli utenti è offerta una pluralità di sistemi, dotati d'intelligenza e gestiti da operatori di servizi di dimensioni molto differenziate.

I *broker* di servizi saranno il canale di più facile accesso e gli unificatori di un'offerta contraddistinta dalla varietà di soggetti, servizi e prezzi.

Gli aggressivi "service provider", in concorrenza tra loro, sceglieranno le vie per sviluppare nuovi servizi anche in funzione del grado di maturità della

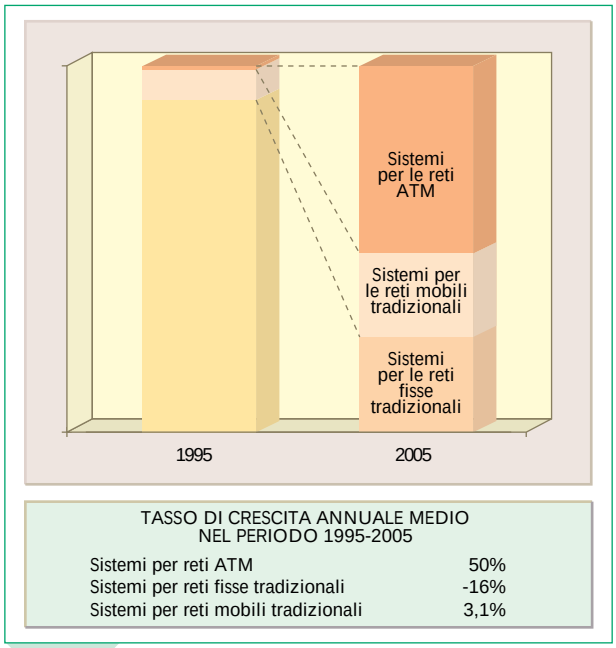


Figura 12 *Variazione negli anni del mercato della commutazione nello scenario Service Mania.*

clientela cui essi si rivolgono e della capacità di spesa ad essa relativa.

I *broker* e i *service provider* sono la vera caratteristica distintiva di questo scenario e i veri motori di sviluppo della Società dell'Informazione, capaci di investire per la nascita dei nuovi servizi.

La rete di accesso è così molto differenziata e dipende dai bisogni di comunicazione dell'utenza e

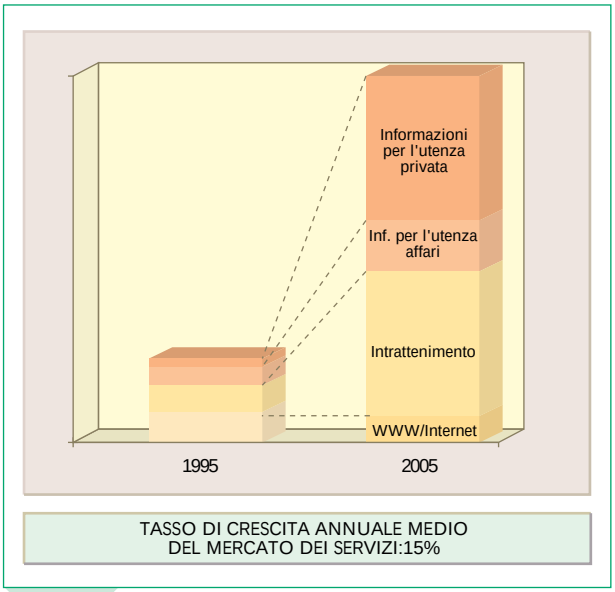


Figura 13 *Mercato dei servizi nello scenario Service Mania.*

dal tipo di servizi richiesti e offerti. Lo sviluppo di reti di accesso a basso costo (quali quelle "wireless") contribuirà a far diffondere i servizi anche in Paesi meno sviluppati.

La rete dati basata sui router si diffonderà rapidamente, in quanto sarà stimolata dalla crescita e dalla varietà dei servizi offerti.

Questo terzo scenario è stato giudicato il più flessibile e forse anche il più auspicabile per uno sviluppo veloce della Società dell'Informazione.

Conclusioni

È evidente che le forze che condizionano gli scenari descritti sono molte e forse assai più numerose di quelle fin qui rapidamente esaminate. Esse poi agiscono in modo differenziato nei diversi Paesi e nei differenti contesti economico sociali.

Il prevalere della tipologia di rete dipende quindi dalla storia delle infrastrutture esistenti in un Paese. Variabili quali: la ricchezza disponibile, la cultura informatica degli utenti, la regolamentazione, la concorrenza sono strettamente correlate tra loro e caratterizzano quel mercato che la tecnologia si sforza di servire.

Al di là delle opportunità aperte dalla tecnologia, a queste forze è lasciata quindi l'ultima risposta per disegnare la fisionomia del futuro nel quale vivremo.

Le telecomunicazioni verso l'assetto numerico

L'evoluzione nei servizi

UMBERTO DE JULIO

Umberto de Julio, Direttore Generale di Telecom Italia, illustra il legame tra le prospettive di sviluppo delle nuove tecnologie e i possibili scenari dell'evoluzione dei servizi.

Desidero anzitutto ringraziare gli amici Francesco Valdoni e Giovanni De Guzzis per avermi invitato a prendere la parola in questa giornata che ha avuto il suo punto centrale nel conferimento della laurea Honoris Causa in Ingegneria al Professore Andrew J. Viterbi, uno studioso che ha contribuito in misura significativa al progresso delle telecomunicazioni e che - come è già stato ricordato dal Professor Roveri - è pervenuto a teorie d'avanguardia per la codifica del segnale in forma numerica.

Tutti noi abbiamo studiato, durante i corsi universitari, l'algoritmo che porta il nome di questo illustre ricercatore ed abbiamo avuto modo di conoscere e di apprezzare l'importanza che il modello matematico da lui elaborato ha avuto negli sviluppi delle comunicazioni via radio negli ultimi trent'anni.

Non nascondo quindi l'emozione che ho provato oggi nell'incontrarlo di persona, quando ho potuto stringergli la mano e apprezzarne la naturale modestia e semplicità, tipiche dei personaggi che hanno contribuito in misura significativa al progresso delle scienze.

Uno scenario che cambia velocemente

Per trattare il tema dell'evoluzione dei servizi e delle prospettive di sviluppo delle nuove tecnologie, oggetto della mia presentazione, come è stato già messo in evidenza dall'Ingegner De Guzzis, ritengo sia fondamentale considerare lo scenario nel quale ci muoviamo, dominato e condizionato da tre elementi guida. Anzitutto la globalizzazione: ogni azienda, in particolare quelle multinazionali o quelle che offrono prodotti o servizi di telecomunicazione, è tesa ad allargare il proprio mercato, e cerca di essere presente in nuovi Paesi, anche lontani da quelli nei quali finora ha tradizionalmente operato.

Una seconda spinta all'evoluzione del settore nel quale operiamo viene dall'innovazione tecnologica. L'importanza, il ruolo e le modalità secondo cui essa oggi è attuata e, in particolare, la flessibilità e la rapidità con cui essa giunge sul mercato, sono temi già approfonditi dall'Ingegner De Guzzis nell'intervento precedente; non ritornerò quindi su considerazioni e su conclusioni già chiaramente esposte, che condivido pienamente, e che confermano l'importanza dell'innovazione tecnologica, per le aziende che operano nel settore dell'*ICT (Information & Communication Technology)*, per adeguare l'of-

ferta alle esigenze del mercato.

Una terza leva di cambiamento infine, oggi presente in misura molto più incisiva che nel passato a noi più vicino, riguarda la convergenza tra telecomunicazioni, tecnologia dell'informazione e media. Una significativa conferma dell'importanza acquisita da questo fattore di crescita è rappresentata dal numero delle fusioni e delle acquisizioni che si sono verificate

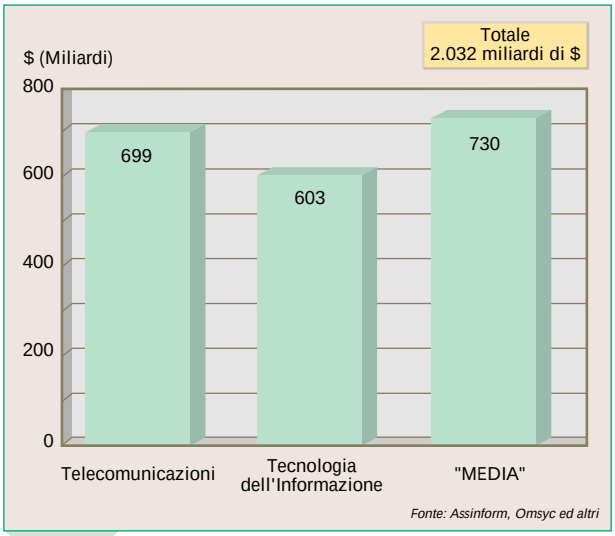


Figura 1 *Il mercato mondiale delle telecomunicazioni, della tecnologia dell'informazione e dei media.*

nel recente passato: secondo la Broadview Associates, nel corso del 1996 si è assistito a un'ulteriore crescita delle fusioni riguardanti i principali settori che convergono verso la multimedialità. Le operazioni censite sono state oltre 3.300 con un valore complessivo di 234 miliardi di dollari (equivalenti a circa 400 mila miliardi di lire), con un incremento di circa il 75 per cento rispetto all'anno precedente.

L'apertura dei mercati alla concorrenza e la convergenza tra i tre settori che ho appena indicato - telecomunicazioni, tecnologia dell'informazione, media - rappresentano assieme fattori di cambiamento degli equilibri esistenti e costituiscono dunque una sfida alle aziende che già operano in questi campi. Ma al tempo stesso questi fattori offrono nuove opportunità di crescita per l'intero comparto.

Come è mostrato in figura 1, stime di settore indicano per il 1996 un valore complessivo del mercato superiore ai 2 mila miliardi di dollari (circa 3 milioni e mezzo di miliardi di lire), un valore quindi pari ad oltre un quarto del *PIL (Prodotto Interno Lordo)* degli Stati Uniti relativo al 1996 (7.342 miliardi di dollari) e quasi doppio del PIL del nostro Paese (pari a 1.208 miliardi di dollari).

Desidero anche sottolineare la tendenza alla crescita del mercato: la convergenza tra i tre settori sopra specificati ha infatti confermato la crescita iniziata già da tempo e che nel 1996 è stata di circa l'8 per cento rispetto all'anno precedente. Dall'esame di questi dati emerge anche un altro elemento significativo: a livello mondiale l'informatica cresce più rapidamente delle telecomunicazioni. Questa tendenza, tuttavia, non trova conferma in Italia dove, come dirò più avanti, le telecomunicazioni si sviluppano con maggiore rapidità dell'informatica.

I principali fattori di cambiamento nell'attuale scenario

Se ora concentriamo la nostra attenzione sui servizi, possiamo cogliere alcuni fattori di cambiamento di rilievo: anzitutto il raggiungimento della saturazione del mercato della *telefonia fissa*, specialmente nei Paesi maggiormente industrializzati; il numero delle connessioni alla rete fissa aumenta ormai in questi Paesi così lentamente da poter considerare quasi costante il numero di utilizzatori collegati alla rete rigidamente.

Possiamo al contrario osservare una crescita estremamente rapida del numero di terminali collegati alla rete cellulare e assistiamo ad un altrettanto elevato sviluppo dei servizi legati alla *mobilità*.

Siamo poi in presenza di altri due fattori di rilievo che influenzeranno la crescita del trasporto dell'informazione e che, insieme alla mobilità - sia pure con effetti diversi - sono destinati a incidere nell'evoluzione dei servizi nei prossimi anni: l'esplosione di *Internet* e l'espansione della *televisione numerica*.

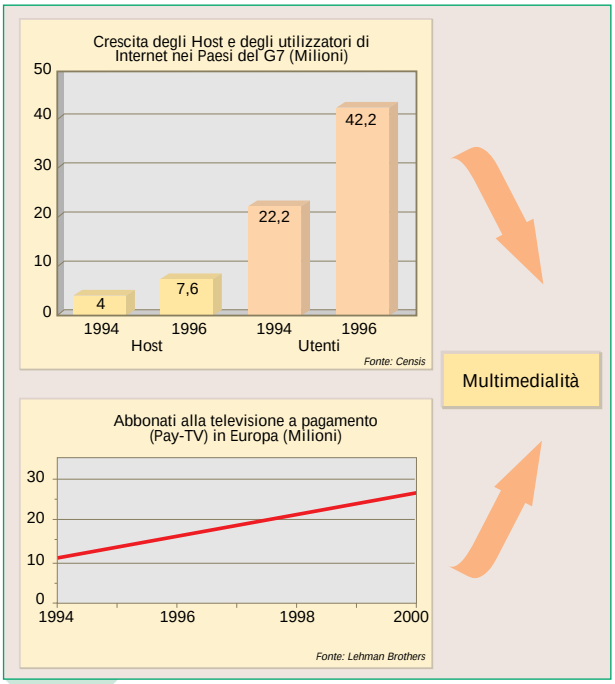


Figura 2 *Crescita di Internet e della televisione numerica: un percorso verso la multimedialità.*

Sono oggi disponibili numerose ipotesi di sviluppo che, pur nelle differenze riscontrabili nelle diverse previsioni, mostrano - tutte - quanto rapido sia stato nei diversi Paesi lo sviluppo dei servizi legati a Internet. In figura 2 sono riportati due dati di consuntivo che mettono in evidenza il raddoppio avuto negli ultimi due anni sia del numero di host collegati a Internet sia del numero degli utilizzatori di Internet.

Gli elementi che abbiamo raccolto finora ci portano a prevedere una forte accelerazione nella crescita di Internet; come è già stato sottolineato dall'Ingegner De Guzzis, i diversi possibili scenari che si presenteranno in futuro avranno un impatto non trascurabile sull'evoluzione del modello di business e sulle modalità con le quali si collegheranno gli utilizzatori di Internet, siano essi i clienti privati o le imprese. Anche il ruolo dei fornitori di servizi nonché di coloro che metteranno a disposizione i contenuti sarà centrale per lo sviluppo di Internet.

L'ultimo filone dei servizi oggi di particolare interesse riguarda l'offerta della televisione numerica: la nuova televisione che, partita dagli Stati Uniti, si sta affermando progressivamente anche in Europa. Le soluzioni adottate differiscono da Paese a Paese a seconda delle diverse situazioni esistenti nell'impiego dei due mezzi trasmissivi: il cavo e il satellite.

Negli Stati Uniti e in numerose nazioni europee, sono già disponibili reti in cavo molto estese. In altri Paesi le reti terrestri sono oggi poco sviluppate o sono praticamente inesistenti, e il collegamento via satellite rappresenta l'unica possibile soluzione per il trasporto a distanza dell'informazione video.

Le prospettive di sviluppo dei servizi ora disponibili vanno dalla ben nota televisione a pagamento (Pay-TV), a servizi a larga banda con un livello di interattività via via sempre maggiore.

Sull'intero scenario che ho fin qui presentato si concentra a mio avviso una scommessa per il futuro. Nessuno di noi riesce a prevedere con una elevata attendibilità quando questi due settori - quello della televisione numerica e quello di Internet - convergeranno verso un sistema che potremmo definire come l'intero insieme dei servizi che si sviluppano su una rete Internet a larga banda.

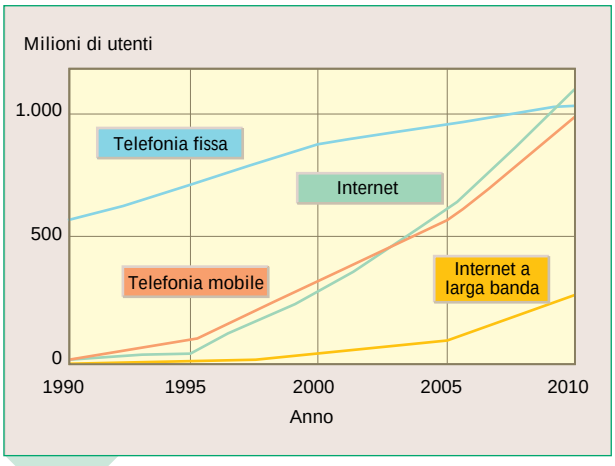


Figura 3 *Lo sviluppo su base mondiale dei servizi.*

La figura 3 riporta il presumibile sviluppo dei servizi nei prossimi anni[1]: la previsione riguarda un arco temporale di circa quindici anni. Al di là dei valori assoluti riportati nel grafico, è opportuno mettere in evidenza alcune chiare tendenze che già cominciano a manifestarsi e che permettono di formulare previsioni sugli orientamenti futuri: entro un certo numero di anni - diciamo fra dieci o quindici anni da oggi - il numero degli utilizzatori di Internet raggiungerà quello dei clienti della rete telefonica tradizionale. Sarà allora difficile distinguere i servizi realizzati secondo modalità tradizionali - legati alla rete fissa esistente - da quelli tradizionali che viaggiano sulla rete Internet.

Nello stesso periodo dovrebbe poi proseguire la rapida crescita del numero di utilizzatori della telefonia mobile che tenderà a raggiungere quello degli utilizzatori della rete fissa.

La percentuale dei clienti dei servizi a larga banda dovrebbe raggiungere, sempre nello stesso arco temporale, un valore compreso tra un quarto o un quinto del totale degli utilizzatori dei servizi di telecomunicazione.

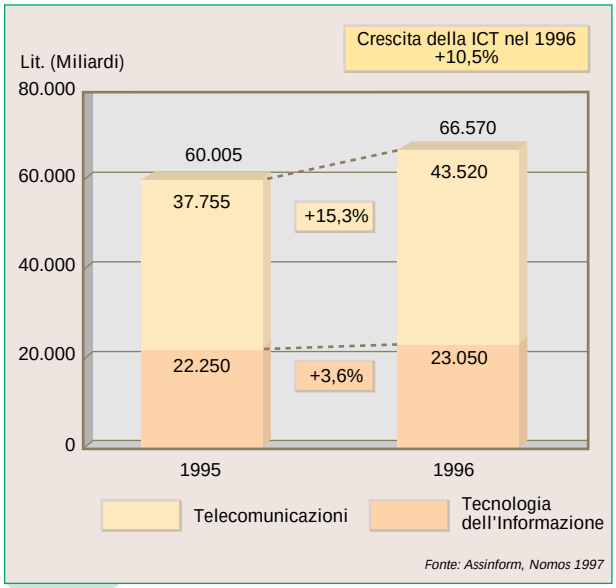


Figura 4 *Il mercato italiano dell'Information & Communication Technology (ICT).*

Scenario futuro in Italia

Possiamo ora cercare di prevedere gli andamenti futuri di crescita per il nostro Paese, sulla base dell'andamento non modesto dei servizi legati all'informatica e alle telecomunicazioni: siamo oggi infatti in presenza di una dinamica ed elevata espansione del settore legato alle telecomunicazioni. Lo confermano i dati del 1996 resi disponibili di recente dall'Osservatorio Assinform, poi ripresi dal Ministero dell'Industria nel rapporto sullo stato dell'informatica e delle telecomunicazioni in Italia: il mercato complessivo del settore è cresciuto di oltre il 10 per cento nel 1996 (figura 4).

(1) *Dècina, M.: Internet e l'Infrastruttura Globale dell'Informazione.* «Notiziario Tecnico Telecom Italia», *Anno 6, n. 2, ottobre 1997.*

Questo risultato è costituito da un aumento del 15,3 per cento delle telecomunicazioni e di un modesto, e forse un po' deludente, 3,6 per cento dell'informatica, che quindi segna ancora il passo rispetto alla crescita rilevabile a livello mondiale, specie nei Paesi più industrializzati. In Italia sono presenti quindi problemi strutturali che è necessario analizzare e risolvere nel prossimo futuro[2].

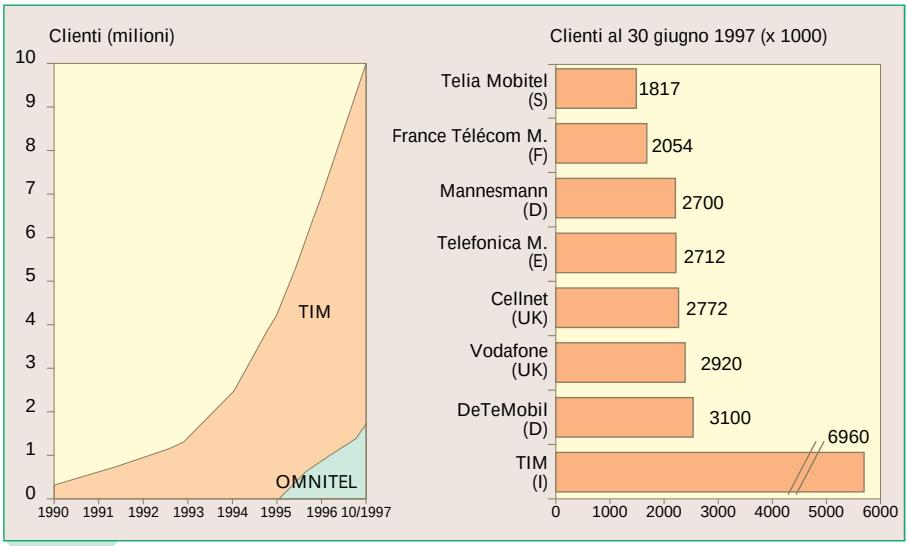


Figura 5 *Confronto tra la mobilità in Italia e quella degli altri Paesi Europei.*

Allo stesso tempo il mercato interno sembra presentare grandi possibilità potenziali di crescita che - sono convinto - sono fortemente favorite dall'apertura dei mercati di telecomunicazione nel 1998 e che per il servizio mobile consentiranno di raggiungere aumenti percentuali esprimibili ancora con numeri a due cifre.

Il numero dei clienti del servizio mobile aveva già superato in Italia alla fine del 1996 la soglia dei 6 milioni (figura 5). Oggi, a metà maggio, abbiamo complessivamente più di 7 milioni di clienti; il nostro Paese è così passato al primo posto in Europa per numero di utilizzatori del servizio mobile, con TIM primo gestore europeo sempre per numero di clienti[3].

Per quel che riguarda Internet abbiamo avuto finora tassi di crescita più ridotti rispetto a quelli rilevati negli Stati Uniti o nei Paesi europei maggiormente industrializzati con i quali siamo soliti confrontarci. Questa crescita più modesta è stata anche un effetto della diffusione più limitata dei personal computer nel nostro Paese.

Per quanto riguarda il futuro, come è mostrato in figura 6, si assisterà probabilmente a circa un raddoppio del numero di clienti ogni anno sia per gli utilizzatori privati che per le imprese. Segnali importanti per lo sviluppo del mercato arrivano anche dalle ultime indagini condotte dall'Osservatorio Alchera che hanno mostrato, nel periodo da ottobre '96 a febbraio '97, una crescita significativa degli utilizzatori di Internet.

Passiamo ora al settore chiamato in maniera un po' generica multimedialità (figura 7): è presumibile che maggiore attenzione sarà rivolta ai servizi offerti ai clienti business, in quanto questi saranno i primi a utilizzare le nuove opportunità di offerta. Essi infatti mettono a loro disposizione una leva per introdurre miglioramenti significativi in termini di efficienza e di efficacia ai propri processi produttivi e, quindi, alla capacità di servire meglio i propri clienti. Il mercato tenderà poi gradualmente ad espandersi fino a interessare il più ampio universo dei clienti privati.

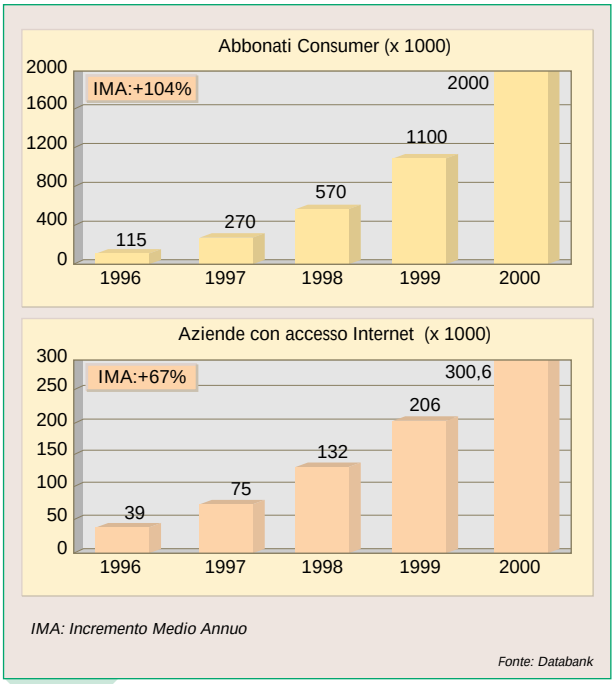


Figura 6 *La crescita di Internet in Italia.*

(2) Un recente studio commissionato dalla Commissione DG XIII di Bruxelles ci indica tuttavia terzi in Europa come pagine Web disponibili, davanti alla Francia.

(3) Nota della Redazione: il sensibile sviluppo previsto dall'Ingegner de Julio si è poi verificato nel corso dell'anno. Quando questo numero della rivista sta per essere mandato alle stampe (dicembre 1997) sono disponibili i dati di consuntivo dell'ottobre scorso che confermano all'Italia la posizione di leader in Europa per il servizio cellulare con oltre 10 milioni di clienti e che posizionano TIM al secondo posto per numero di clienti, tra i gestori mondiali (dopo l'operatore giapponese NTT DoCoMo), con oltre 8,2 milioni; si prevede di superare in Italia la soglia di 11 milioni di utenti mobili a fine 1997 e di raggiungere i 9 milioni di clienti per TIM.

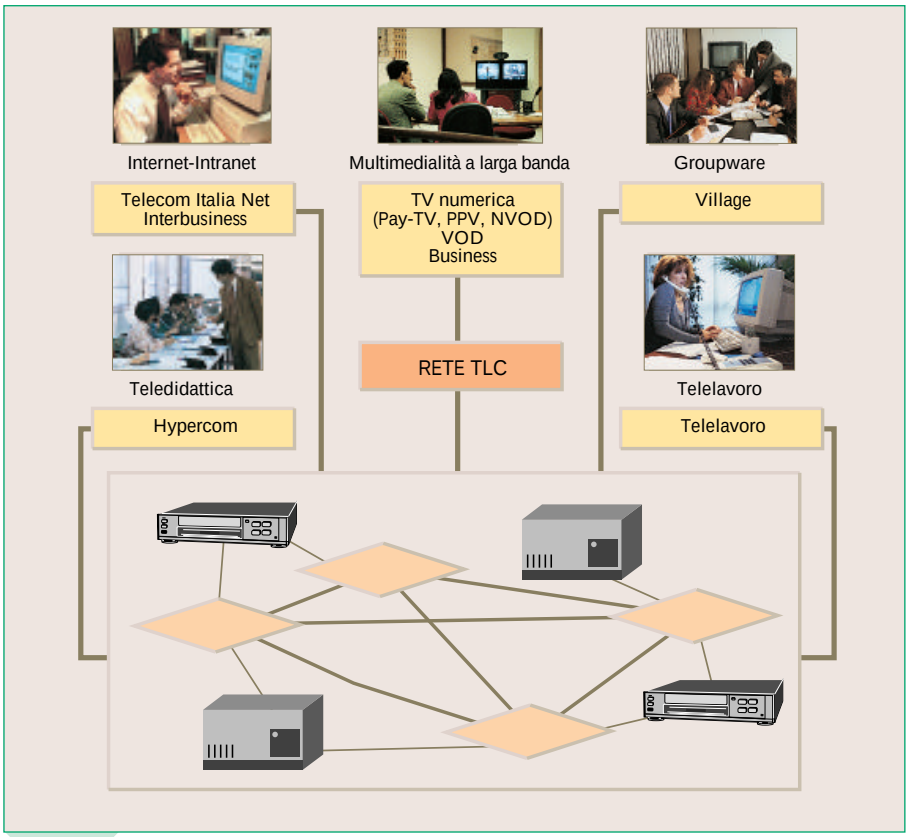


Figura 7 *Il cammino di Telecom Italia verso la multimedialità.*

Nel quadro fin qui presentato è possibile cogliere un'evoluzione significativa dell'offerta di Telecom Italia caratterizzata da un livello di innovazione sempre più spinto.

Come aumentare la competitività del Paese

Vorrei soffermarmi nell'ultima parte del mio intervento su alcune personali riflessioni e su alcune proposte che riguardano alcuni potenziali strumenti per rendere il Sistema Paese sempre più competitivo.
Ritengo infatti che sia opportuno individuare e mettere a punto nuovi modelli di sviluppo che di fatto stimolino la creatività dei singoli, favoriscano la nascita e il consolidarsi di nuova imprenditorialità e creino dunque innovazione.
Ma per perseguire queste finalità - diversamente dal passato - non è più sufficiente incoraggiare la ricerca svolta dai maggiori complessi industriali, in quanto l'innovazione non è più appannaggio quasi esclusivo dei grandi gruppi manifatturieri ma è prodotta in maniera diffusa.
Una conferma la ritroviamo negli ultimi sviluppi di Internet: in quest'area una "miriade" di piccole imprese realizzano innovazione sia nei prodotti che nei servizi. La dinamica di crescita sul mercato di queste imprese è notevole. Tra gli esempi di maggior rilievo potrei citare quelli di *Netscape* (WEB browser), *Pointcast* (push delivery), *WEB TV* (Internet non legato ai personal computer) e *SUN* (Java; Network Computer).

Da un punto di vista generale, ritengo poi che, per creare condizioni che favoriscano innovazione nei settori tecnologicamente più avanzati, convenga seguire in Italia modelli di incentivazione e di accelerazione analoghi a quelli utilizzati in altri Paesi.

Dobbiamo perciò riflettere sulla opportunità - o meglio sulla necessità - di dare maggiore fiducia

Da sinistra: Aldo Roveri, Andrew J. Viterbi assieme alla moglie Signora Erna, Guido Vannucchi, Francesco Valdoni e Giovanni De Guzzis seguono l'intervento di Umberto de Julio.

all'innovazione che presenta un più elevato grado di rischio ma che può rappresentare l'elemento trainante per lo sviluppo di nuovi servizi.

È perciò opportuno l'avvio di un ciclo virtuoso che, partendo dalle idee e attribuendo a ciascuna di esse risorse finanziarie adeguate, permetta di trasformare le proposte in prodotti o in servizi e consenta, quindi, di ottenere così un ritorno degli investimenti.

Silicon Valley: un esempio da studiare

Un modello che potrebbe essere adottato per lo sviluppo dell'innovazione è mostrato in figura 8: i dati riportati si riferiscono ad un'area degli Stati Uniti, la Silicon Valley, dove forse sono stati ottenuti risultati tra i più significativi nei settori nei quali

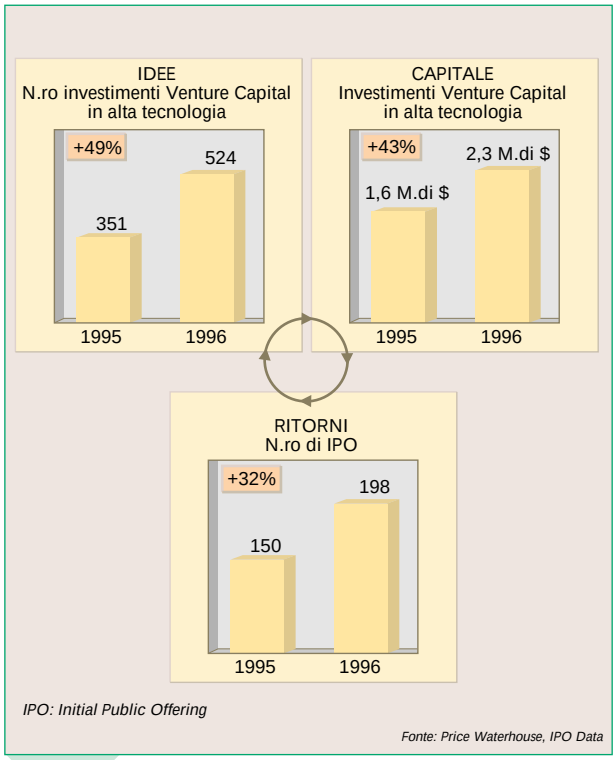


Figura 8 *Il caso Silicon Valley: un modello di innovazione.*

operiamo: in quest'area il numero di *Venture Capital* (capitale a rischio) - cioè di investimenti con prospettive di reddito e rischio elevati, operato da finanziarie verso aziende di norma con dimensioni modeste ma tecnologicamente molto avanzate - è cresciuto dal 1995 al 1996 di circa il 50 per cento; mentre gli investimenti sono aumentati nello stesso periodo di oltre il 40 per cento, raggiungendo il valore di 2.300 miliardi di dollari (4 milioni di miliardi di lire circa).

Anche il numero dei ritorni, cioè dei successi ottenuti con questi investimenti, è cresciuto - sempre nello stesso periodo - di circa il 32 per cento.

È anche aumentato il numero di aziende che, finanziate con Venture Capital, hanno assunto una dimensione che ha loro permesso di entrare nel mercato borsistico attraverso i cosiddetti *IPO (Initial Public Offering)* e che cominciano a giocare così un ruolo di Public Company.

Sempre rimanendo nell'area geografica della Silicon Valley, possiamo osservare che l'attenzione maggiore per gli investimenti a rischio è volta a

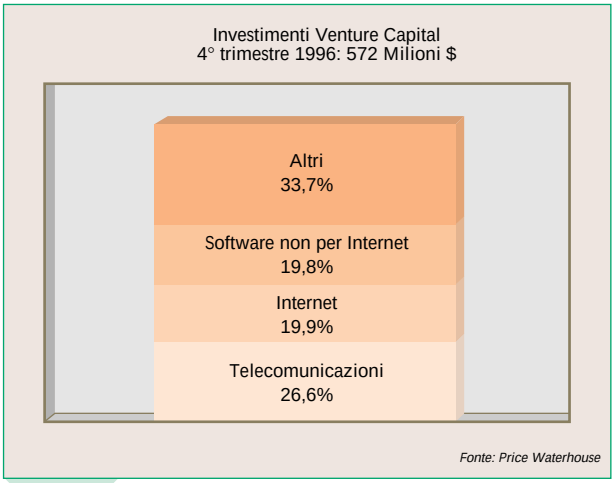


Figura 9 *Il caso Silicon Valley: la suddivisione percentuale degli investimenti in Venture Capital nei settori avanzati.*

ricerca applicata nei campi contigui delle telecomunicazioni e dell'informatica: ad essa è stato infatti indirizzato il 67 per cento di questi investimenti, suddivisi tra un 20 relativo a Internet, un 27 alle telecomunicazioni e un 20 per cento a software non per Internet (figura 9).

Anche altri settori di ricerca avanzata, ad esempio quelli relativi alle biotecnologie, sono considerati come opportunità per lo sviluppo di nuovi prodotti di rilevante interesse e, quindi, potenzialmente idonei a garantire il ritorno degli investimenti.

La figura 10 mostra infatti come è stato investito il Venture Capital nel quarto trimestre del 1996: dei 572 milioni di dollari (1.000 miliardi circa di lire), il 20 per cento è stato indirizzato a iniziative *seed*, cioè a ricerche in settori altamente innovativi. E sembra importante rilevare che Internet è divenuto il protagonista dell'innovazione: ad essa sono stati assegnati il 32 per cento degli investimenti iniziali ad alto rischio (di tipo seed), superando così il valore di quelli indirizzati alle telecomunicazioni, che risultano ora di poco maggiori del 25 per cento.

Il contributo di Telecom Italia

Anche STET assieme a Telecom Italia[4] ha colto l'importanza dell'innovazione e ha avviato una serie di iniziative di rilievo per cercare di creare le condi-

(4) Nota della Redazione: al momento di questa presentazione non era ancora avvenuta la fusione tra STET e Telecom Italia.

zioni necessarie a stimolare lo sviluppo di prodotti e di servizi avanzati e di diffondere la cultura multimediale in Italia.

In questo quadro si inseriscono le iniziative dei Cantieri Multimediali, che creano le condizioni favorevoli per lo sviluppo di nuove idee e servizi fin dai primi passi dalla ricerca svolta presso l'Università; degli incubatori, che permettono la realizzazione e la verifica delle nuove idee congiunta all'imprenditoria giovanile in modo da concretizzare il progetto industriale e da risolvere le difficoltà burocratiche e operative all'avvio dell'iniziativa che, se appesantita, rischierebbe di far fallire l'innovazione. A queste iniziative si affiancano il Venture Capital e Fintech[5], che mettono a disposizione capitali di rischio per finanziare iniziative che presentano un elevato carattere innovativo e potenzialità di successo. Si è cercato così di ricalcare in Italia i modelli di sviluppo, cui prima facevo cenno, seguiti negli Stati Uniti; e, quindi, di stimolare, o almeno di favorire, i rapporti tra il mondo della ricerca e quello delle imprese e, d'altra parte, di ridurre il tempo tra la disponibilità dei risultati della ricerca e l'approntamento dei prodotti o dei servizi sul mercato.

Una materia tutta ancora da approfondire

Avviandomi a concludere, desidero sottolineare che in questo periodo sono presenti due fattori trainanti che offrono ai settori nei quali operiamo un'opportunità di sviluppo molto significativa: il primo riguarda il processo di convergenza che, rimescolando imprese e mercato, può costituire, come ho già avuto modo in precedenza di osservare, una premessa importante per la crescita complessiva del nostro settore.

Il secondo fattore è legato al processo di liberalizzazione delle telecomunicazioni, ora in corso nella Comunità Europea, e che rappresenta sicuramente un elemento di sviluppo sia per i nuovi entranti sia per i gestori che già operano in quest'area.

Le imprese, a mio avviso, stanno già agendo attivamente per cercare di cogliere queste nuove opportunità di crescita. Approfitto della presenza dell'Ingegner Vannucchi, che parlerà dopo di me, per citare, a titolo di esempio, un accordo recente tra RAI e STET per la creazione di una piattaforma unica, atta a favorire lo sviluppo dei servizi televisivi numerici in Italia. Questo accordo costituisce un esempio di grosso rilievo sulle possibili convergenze tra le tecnologie impiegate dai diversi settori, e mostra quali possono essere le occasioni di sviluppo che si presenteranno nel prossimo futuro. Penso, tuttavia, che non sia sufficiente l'impegno solo delle imprese e, in particolare, di quelle maggiori o un collegamento stretto tra queste e il settore della ricerca applicata.

Ritengo, infatti, che il processo di innovazione

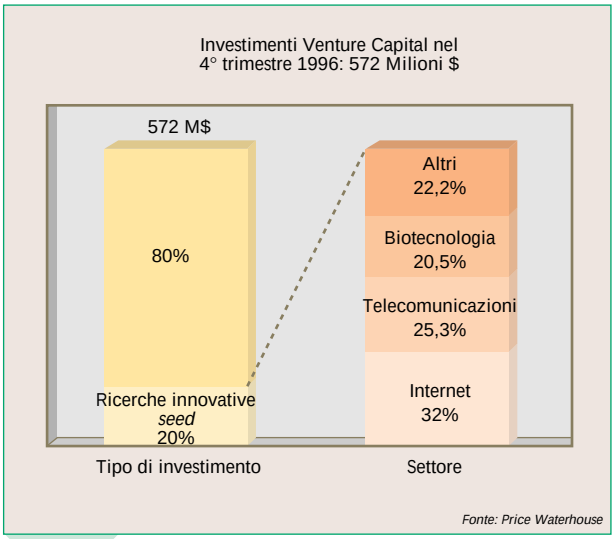


Figura 10 *Il caso Silicon Valley: settori avanzati interessati dagli investimenti Venture Capital.*

debba coinvolgere l'intero Paese e debba assicurare certezza agli investimenti attraverso leggi e regolamenti. L'esame finale, ad esempio, del nuovo regolamento delle telecomunicazioni, in corso in queste settimane, consentirà di fare un passo significativo in avanti a questo scopo[6].

Anche il mondo della finanza deve acquisire fiducia nell'investire in nuove idee valutando i rischi ma anche i benefici che possono essere colti da iniziative promettenti in campi di sicuro interesse; ma allo stesso tempo deve mostrarsi sempre più flessibile in modo da sapersi adattare ad uno scenario in continuo cambiamento.

Un'attenzione particolare va rivolta, quindi, verso l'innovazione con la convinzione che attraverso questo lavoro congiunto dei diversi attori coinvolti nel settore si realizzi lo sviluppo di nuovi servizi e quindi un potenziamento ed un allargamento del mercato. Questi sono elementi - tutti - che concorrono sicuramente a creare nuove opportunità di crescita del nostro Paese e a mantenerlo al livello di quelli più innovativi. Il processo ora proposto potrà infatti avere ricadute importanti anche sull'intera economia nazionale e potrebbe avere un'influenza positiva sui livelli di occupazione.

Converrà quindi continuare ad approfondire tutti assieme e a decidere con rapidità perché i tempi per i cambiamenti si sono di molto abbreviati rispetto al passato, anche rispetto a quello a noi più vicino, e si mantiene la competitività sul mercato solo tenendo il passo con l'accelerazione tecnologica presente a livello mondiale.

Ringrazio per l'attenzione.

(5) *Società recentemente formata da STET e Mediocredito Centrale.*

(6) *Nota della Redazione: dopo qualche mese da questo intervento sono stati approvati il "Regolamento per l'attuazione di direttive comunitarie nel settore delle telecomunicazioni" DPR n. 318 del 19 settembre 1997 e l'"Istituzione dell'Autorità per le garanzie nelle comunicazioni e norme sui sistemi delle telecomunicazioni e radiotelevisivo" legge 249 del 31 luglio 1997.*

Le telecomunicazioni verso l'assetto numerico

Verso la Società dell'Informazione: opportunità e rischi

GUIDO VANNUCCHI

Guido Vannucchi, Vice Direttore Generale della RAI, indica le opportunità e i rischi nel cammino verso la Società dell'Informazione.

Ringrazio anch'io Francesco Valdoni per l'opportunità che mi ha dato con questo intervento.

Ho accettato l'invito con entusiasmo, anche se avevo avuto la fortuna di anticipare lo scorso dicembre quell'emozione, cui ha già accennato Umberto de Julio, alla Columbia University dove ho avuto il piacere di conoscere il Professor Viterbi in un simposio che chiudeva un ciclo delle celebrazioni in onore di Marconi. Insieme ad Andrew J. Viterbi ho potuto incontrare anche il dottor Lucky. Certamente un'accoppiata di nomi mitici per gli esperti in trasmissione oggi presenti! E lo scenario era completato dai tre padri di Internet: Robert Kahn, Vinton G. Cerf e Leonard Kleinrock.

La Società dell'Informazione

Per ragioni alfabetiche sono quasi sempre l'ultimo a parlare; il taglio del mio intervento è volutamente un po' diverso: dopo un cenno all'evoluzione delle sole tecnologie televisive mi concentrerò, in particolare, sulle possibili conseguenze sociali della Società dell'Informazione. Ritengo infatti che anche scienziati e tecnologi, pur entusiasmandosi per l'evoluzione della tecnica, debbano sempre prestare molta attenzione alle conseguenze che determinate tecnologie possono portare.

A questo scopo ho riesaminato i miei interventi anche molto lontani nel tempo cercando di capire quanto è ancora valido e quanto ormai risulta superato. Parto perciò da molto lontano esaminando le diverse fasi storiche della Società civile, della Società preindustriale e di quella dell'Informazione o Società postindustriale (figura 1). Per inciso ricordo che Società dell'Informazione, Società del terziario avanzato o Società postindustriale sono tutte denominazioni che praticamente si equivalgono.

Le sofferenze che stiamo vivendo in questo momento, in particolare l'aumento della disoccupazione, possono essere tipici elementi che caratterizzano un momento di *transizione*: tutte le volte infatti che nella storia si è passati da uno stadio di civiltà ad uno successivo (ad esempio dalla civiltà agricola a quella industriale), si sono sempre presentate perturbazioni negative che solo successivamente sono state riassorbite da nuovi sostanziali vantaggi.

Per verificare la veridicità di queste affermazioni occorre tuttavia attendere di passare dalla Società Industriale a quella dell'Informazione. Un economista americano, premio Nobel nel 1981, James Tobin, osservava infatti che finora la Società dell'Informazione è stata molto deludente nei riguardi dell'aumento di produttività complessiva del sistema: in particolare è stato ottenuto un aumento a livello di singole fasi, ma non un aumento significativo a livello di processo visto nel suo insieme.

A questo proposito si può riportare un esempio

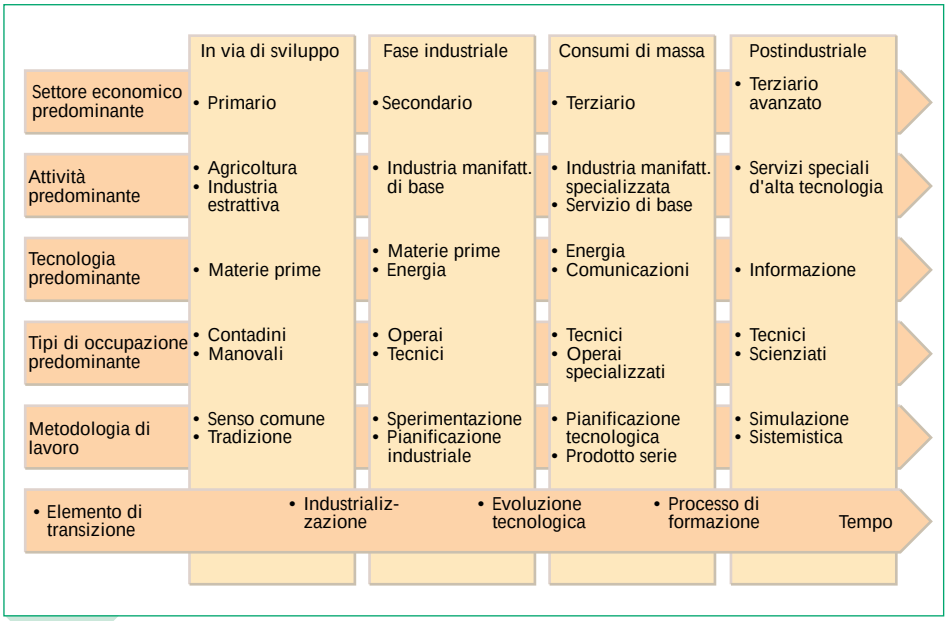


Figura 1 *Alcune caratteristiche tipiche delle strutture sociali riferite all'evoluzione delle civiltà.*

banale: tutti noi ricordiamo che venti anni fa si diceva: "adesso che stanno per arrivare gli elaboratori e si introdurrà l'Office Automation tutta la carta che oggi ci affligge sparirà completamente". Mai vista tanta carta come in questo momento!

Tornando alle transizioni, vorrei ricordare che la *popolazione mondiale è variata a gradini* perché, pur essendo limitate le risorse a disposizione dell'umanità, in ciascuna transizione è cambiato in misura significativa il livello di produttività conseguente ai livelli di conoscenza tecnologica.

In particolare, prima della rivoluzione agricola, il mondo era assestato sui 10 milioni di abitanti; dopo la rivoluzione agricola esso si è portato a 700 milioni di abitanti e si è fermato per molti anni su questo valore. Con la rivoluzione industriale si è passati dai 700 milioni di abitanti ai 5 miliardi attuali: tutti voi tuttavia conoscete bene quali furono le perturbazioni sociali, specie quelle conseguenti all'introduzione dei telai meccanici in Inghilterra alla fine del Settecento, che hanno caratterizzato il passaggio dalla civiltà agricola a quella industriale.

La civiltà industriale è stata essenzialmente caratterizzata da un fattore di grande rilevanza: *la produzione di grandi quantità di energia a costi contenuti*. Altro elemento fondamentale, che ha caratterizzato la civiltà industriale, è stato l'aumento significativo di produttività del lavoro. La transizione ha infine causato il decadimento delle due classi sociali - la nobiltà e il clero - che avevano caratterizzato il periodo precedente, sostituite con la borghesia e il proletariato.

Tutti questi cambiamenti sono stati realizzati attraverso *crisi economiche ricorrenti, ristrutturazioni e disoccupazione*. L'enorme aumento di produttività per l'epoca ha anche caratterizzato una ricerca sempre più spasmodica di nuovi mercati fuori dal proprio territorio. Un aspetto negativo, di cui soffriamo tuttora, riguarda infine una serie di effetti dannosi sull'ambiente causati dalla civiltà industriale.

La nuova Società postindustriale (ovvero la Società dell'Informazione), è quella che oggi cominciamo a vivere ed è caratterizzata, in particolare, da un intreccio di diverse tecnologie: le telecomunicazioni, l'informatica e la microelettronica; il nucleare; la tecnologia spaziale; la robotica (ancora nelle fasi iniziali); la biotecnologia.

Il fenomeno peculiare che ha caratterizzato questo *intreccio di diverse tecnologie* è stato la forte *fertilizzazione incrociata ("cross fertilization")* in cui il ruolo di locomotiva è stato assunto dall'*ICT (Information & Communication Technology)*.

È stato anche avviato, ormai da diversi anni, un processo di *dematerializzazione dei beni* che, fondamentalmente, è caratterizzato dal passaggio dai prodotti semplici a quelli sempre più sofisticati in cui il contenuto di integrazione di "Software" è andato via via crescendo (figura 2). Dal punto di vista del consumatore si è avuta, contemporaneamente, una sensibile riduzione nell'impiego di materia prima, nell'energia necessaria alla produzione e, in ultima analisi, nei costi dei prodotti.

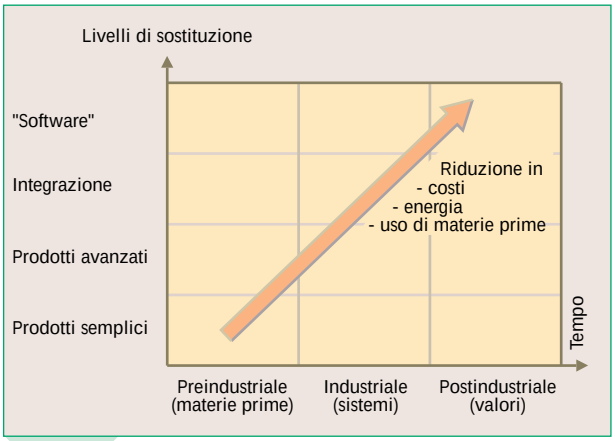


Figura 2 *La dematerializzazione dei beni.*

Nel corso della nostra esperienza di lavoro abbiamo avuto modo di assistere alla trasformazione di un apparato di telecomunicazioni da prodotto "hardware", ricco di valvole, condensatori, resistenze e altri componenti, ad un insieme costituito da un modesto numero di circuiti ad alta integrazione. Non è tuttavia sparita l'intelligenza di cui è stato necessario disporre,

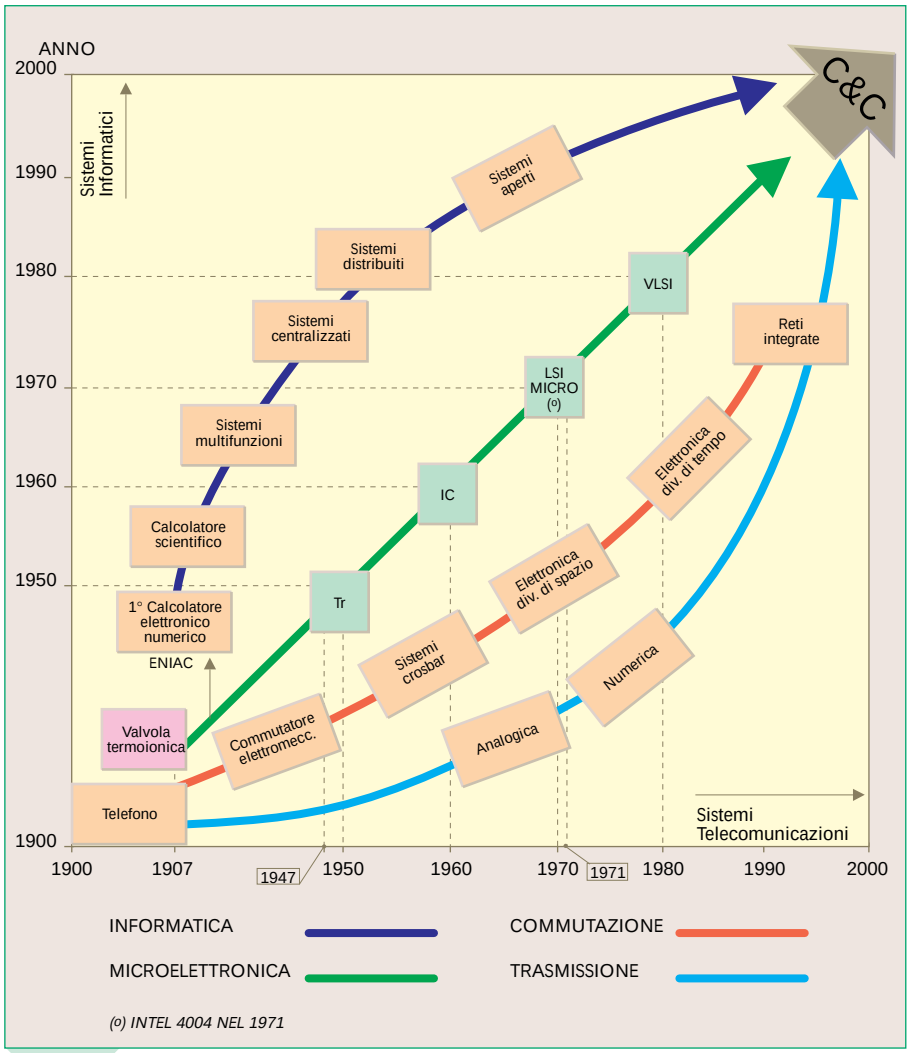


Figura 3 *Interazione e confluenza fra telecomunicazioni e informatica in funzione dell'evoluzione tecnologica.*

a monte, per concepire queste soluzioni moderne; essa anzi è aumentata in misura sensibile.

La convergenza tra l'informatica e le telecomunicazioni

Da quanto ho detto sulla Società dell'Informazione, la *convergenza tra informatica e telecomunicazioni* è risultata essere uno dei cardini fondamentali dell'intero processo.

Per molti anni si è vissuti tuttavia nell'illusione che questa convergenza fosse essenzialmente caratterizzata da un processo tecnologico, senza che essa avesse conseguenze immediate sul mercato. In particolare, negli anni Ottanta, si avviarono processi di fusione tra le aziende informatiche e quelle di telecomunicazioni, falliti successivamente proprio perché prematuri o in quanto non erano stati integrati i rispettivi mercati.

La microelettronica è stata assolutamente un fattore fondamentale per questa convergenza e lo è tuttora. Un altro passo di grande rilievo, caratteristico degli ultimi anni, è stata l'introduzione di standard universali di comunicazione e lo sviluppo dei protocolli di comunicazione di tipo Internet (TCP-IP).

Vorrei mostrare un diagramma molto noto (figura 3) a cui sono particolarmente legato (nelle mie presentazioni ho avuto occasione di mostrarlo numerose volte): è uno strano diagramma che porta in ordinate e in ascisse gli anni. La rappresentazione caratterizza le tappe fondamentali dello sviluppo, rispettivamente dell'informatica e delle telecomunicazioni: quest'ultimo sviluppo a sua volta è suddiviso in due tecnologie distinte, la trasmissione e la commutazione, che solo con l'introduzione della commutazione a divisione di tempo sono confluite nella *comune tecnologia delle telecomunicazioni numeriche.*

L'asse portante del processo di convergenza è la retta centrale che scandisce le tappe fondamentali della tecnologia dei semiconduttori e i progressi dell'integrazione. È abbastanza interessante rilevare che l'informatica, all'inizio del processo, è stata debitrice alle telecomunicazioni per la componentistica, a cominciare dal transistor fino al progetto di computer ad alta affidabilità. L'informatica ha poi "restituito" alle telecomunicazioni molti dispositivi di tecnologia raffinata, quali il microprocessore.

Questo diagramma, presentato per la prima volta negli anni Settanta da Kobayshi, Presidente della NEC, ipotizzava una convergenza anticipata rispetto a quanto poi in realtà è avvenuto. Forse Kobayshi si basava sulla valutazione (rivelatasi poi errata), che la convergenza tecnologica e quella di mercato arrivassero assieme.

Lo scenario delle quattro C

Più recentemente, in luogo della dizione *ICT (Information & Communication Technology)* si preferisce parlare dello *scenario delle quattro C.*

Che cosa sono le quattro *C*? Alle *Communication* (cioè le Telecomunicazioni), e al *Computing* (Informatica) si aggiungono altre due *C*, e precisamente il *Content* (Contenuto), e il *Consumer*, ossia la capacità di produrre a prezzi contenuti prodotti altamente sofisticati per l'utente finale. Un decodificatore televisivo numerico, ad esempio, è un terminale estremamente sofisticato che deve avere un prezzo molto contenuto;

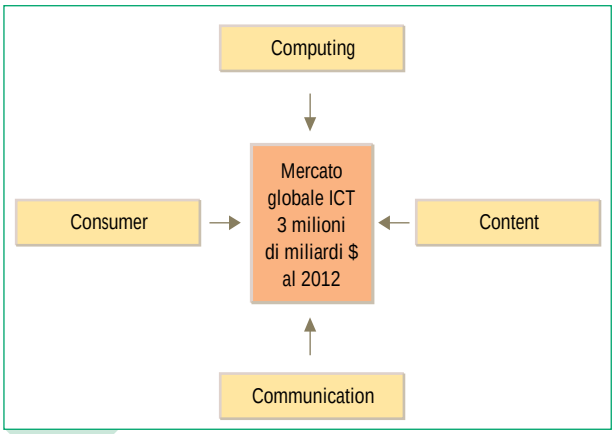


Figura 4 *Il mercato nel 2012 legato allo scenario delle quattro C.*

questo obiettivo è perseguibile solo con componenti ad integrazione molto elevata e con produzioni altamente automatizzate.

L'insieme delle quattro C (figura 4) rappresenterà nel 2012, in base alle previsioni, un mercato di 3.000 miliardi di dollari, valore così elevato che rinuncio a trasformarlo in lire.

È anche abbastanza interessante l'esame di un diagramma (figura 5) che mostra in funzione degli anni il tasso di aumento percentuale delle grandi direttrici di sviluppo che hanno negli ultimi anni segnato il mercato ICT. All'inizio è cresciuto in misura maggiore il settore delle "Reti Corporate"; successivamente è toccato all'informatica distribuita per il grande successo dei personal computer. Oggi il problema di fondo è

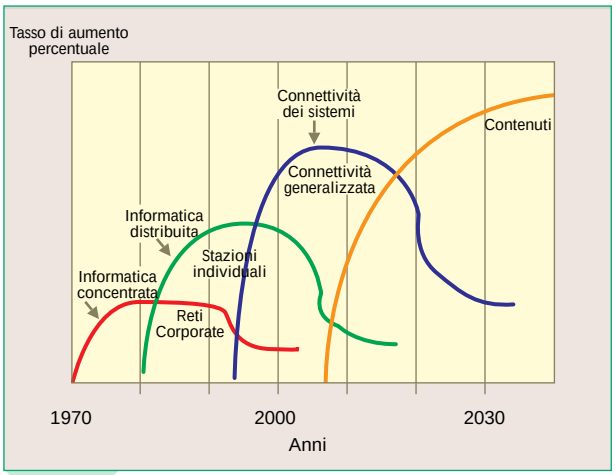


Figura 5 *Evoluzione tra il 1970 e il 2030 dell'ICT (Information & Communication Technology).*

quello della "connettività" generalizzata, e quindi di un settore che aumenta con un ritmo di crescita molto elevato. Negli anni futuri, intorno al 2005-2008, è previsto che il *Content* (Contenuto) rappresenti il settore a più alto tasso di crescita che raggiunge, nella rappresentazione sopra menzionata, una porzione di gran lunga maggiore rispetto agli altri tre settori.

Evoluzione della multimedialità

Un fenomeno inaspettato nell'evoluzione della multimedialità è stato quello mostrato nel diagramma di figura 6: con esso l'autore si è "divertito" a calcolare l'ammontare di elementi binari di informazione trasmessi in un anno per i differenti tipi di informazioni. Questo spiega perché i numeri in ordinate siano stratosferici!

Dal diagramma risulta che la "Telefonia" è cresciuta a un tasso relativamente costante; la "Trasmissione Dati" (che si preconizzava dovesse superare la telefonia) rimane due ordini di grandezza inferiori (anche per la maggiore efficienza raggiunta nella stessa trasmissione dati); infine i "documenti scritti" (nonostante tra questi sia considerato anche il facsimile) sono cresciuti in misura alquanto modesta.

Un fenomeno del tutto inaspettato è stato invece quello *della crescita di informazioni rappresentate da immagini in movimento* (televisione) che, negli ultimi anni, sono cresciute a un ritmo estremamente sensibile superando, come capacità d'informazione, la

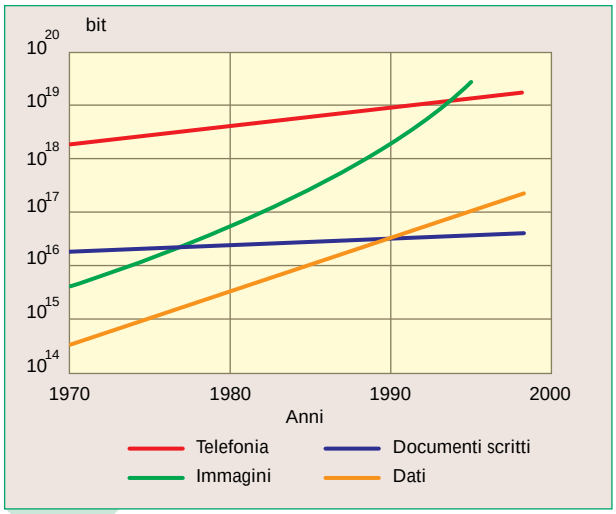


Figura 6 *Numero di elementi binari d'informazione (bit) trasmessi in un anno per differenti tipi di informazioni.*

stessa telefonia. Questa affermazione non desterebbe naturalmente alcuna meraviglia se la curva ad essa relativa rappresentasse le immagini "ricevute" in quanto i televisori sono oggi in numero pressoché uguale a quello dei telefoni e il segnale televisivo ha un contenuto in bit assai più alto della voce. Il diagramma intende invece rappresentare il numero di informazioni visive *differenti alla fonte* e il risultato mostrato presenta perciò un andamento del tutto inatteso.

Gli sviluppi tecnologici che hanno caratterizzato l'esplosione delle trasmissioni visive sono: la *numerizzazione* dei segnali televisivi e la conseguente facilità di *compressione* degli stessi segnali con l'eliminazione delle ridondanze. In particolare la definizione dello standard universale di codifica numerica televisiva MPEG ha dato una notevole accelerazione

a questo processo di compressione. (Va tuttavia ricordato che, storicamente, la compressione è nata essenzialmente per rendere possibile la trasmissione di *segnali ad alta definizione*).

Ritengo ora utile passare brevemente in rassegna *i moderni sistemi di distribuzione dell'informazione video.*

I *satelliti numerici* si sono affermati con maggiore facilità perché la struttura del mezzo è rimasta sostanzialmente inalterata rispetto al precedente impiego con segnali analogici. Un "trasponder", nato per un canale televisivo analogico, è però oggi in grado di trasportare da 7 a 8 canali televisivi numerici adottando come standard di modulazione il semplice sistema 4PSK (il tubo a onda progressiva con le sue non linearità non permette modulazioni più sofisticate).

Mediante una codifica dei segnali con il metodo di Viterbi e l'adozione anche dei codici correttori Reed-Salomon, si riescono oggi ad ottenere prestazioni estremamente avanzate anche con satelliti non di altissima potenza e con antenne di dimensioni molto modeste.

Nell'ambito della standardizzazione internazionale è stata anche resa facilmente possibile l'*interconnessione tra satellite e cavo* adottando per quest'ultimo una modulazione 64 QAM ed eliminando la codifica di Viterbi (in quanto superflua poiché il mezzo trasmissivo è privo di "fading").

Con questi due accorgimenti l'intero multiplo di 7 o di 8 canali allocati in un trasponder satellitare (complessivamente equivalente ad una velocità di cifra, bitrate, di 32 Mbit/s) può essere allocato in una "slot" analogica di 8 MHz, che, nel cavo, corrisponde all'intervallo di frequenza riservato ad un canale analogico.

Un altro *sistema con caratteristiche straordinarie è l'ADSL* che permette di impiegare lo stesso doppino di utente per trasmettere uno o due canali televisivi.

La complessità della realizzazione dei circuiti integrati per ADSL (dell'ordine di 1÷2 milioni di componenti) ha rallentato l'introduzione sul mercato di questi sistemi, ma quando i componenti integrati saranno disponibili, il sistema avrà indubbiamente un grande rilancio.

Un altro possibile metodo di diffusione è quello della televisione numerica a diffusione terrestre che si fonda, oltre che sulla compressione dei segnali televisivi, su una tecnica di modulazione estremamente innovativa *(modulazione OFDM)*, in grado di dare risultati molto prossimi ai limiti di Shannon, pur avendo notevoli gradi di flessibilità ed una elevata resistenza ai disturbi. È veramente una tecnica di modulazione straordinaria!

Vanno infine ricordati anche i *sistemi cellulari a microonde ad altissima frequenza*, che operano intorno ai 40 GHz e che presentano caratteristiche tecniche sostanzialmente simili a quelle dei satelliti. Obiettivo di questi sistemi è la competizione con i cavi; essi forniscono perciò anche un canale di ritorno nell'ambito della stessa ampia banda impiegata.

La multimedialità interattiva personalizzata

A questo punto è utile fare una considerazione: si parla oggi di satelliti con uso interattivo (e sicuramente essi avranno un successo notevole per reti Intranet a larga banda). Tuttavia più in generale, se si pensa ai satelliti per uso di "video on demand" (con richiesta fatta tramite la linea telefonica) occorre tener presente che, per quanto tecnicamente possibile, una frequenza da satellite è un bene troppo prezioso per essere dedicato a una sola persona, ed in ogni caso il suo impiego non risulterebbe economico.

L'applicazione della *multimedialità interattiva personalizzata* trova pertanto il suo campo di utilizzo migliore con reti del tipo in cavo a struttura "stellare" (non ad "albero").

Per concludere questa parentesi tecnica vorrei ricordare l'importanza strategica del *set top box*, ossia

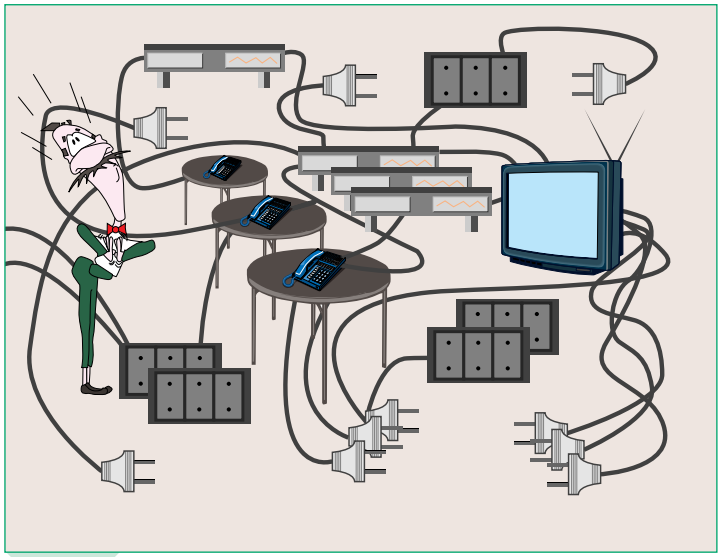


Figura 7 *Senza precisi standard ci prepariamo forse ad una gran confusione?*

del *decodificatore* ("decoder") video numerico connesso al normale apparecchio televisivo. Il decodificatore può ricevere i programmi dal satellite ovvero dal cavo od ancora dal doppino d'utente (ADSL). Nel prossimo futuro sarà certamente indispensabile disporre di un apparato unificato in cui, con piccoli moduli intercambiabili (o già tutti contenuti nel decodificatore), si possano soddisfare le diverse esigenze sopra menzionate: come analogia, si pensi che oggi non fa meraviglia che un televisore sia in grado di ricevere tutti gli standard televisivi analogici mentre dieci anni fa questo obiettivo appariva una follia economica.

Un elemento fondamentale del decodificatore che accresce ulteriormente la confusione è il tipo di "accesso condizionato", che si riflette nel sistema di criptaggio, elemento cardine delle televisioni a pagamento (Pay-TV).

Si vanno affermando al riguardo due tendenze: la prima è rappresentata dai tipi di decodificatori proposti da "service provider" che intendono fare

dell'apparato un elemento proprietario (e quindi che desiderano difendere, attraverso la tecnologia, anche il loro mercato); l'altra tendenza privilegia i decodificatori universali nei quali si sostituisce solo un piccolo modulo denominato "Condition Access Module" che ha all'interno il circuito di decriptaggio in grado di caratterizzare un particolare "service provider".

Spero che la tendenza verso forme ad accesso condizionato aperto si rafforzi in quanto ritengo che questo metodo rappresenti un elemento essenziale di libertà di mercato.

Tenendo conto di tutte le diverse opzioni, se non si arriva presto ad un grado di standardizzazione molto spinto, la situazione della terminazione di utente può rimanere assai confusa (figura 7).

La rivoluzione multimediale e l'avvento della Società dell'Informazione

A questo punto è opportuno ricordare cosa si deve intendere per *multimedialità*: essa è l'*insieme delle diverse forme*, ossia testi, grafici, fonia, immagini (fisse o in movimento), tra loro comunque combinati, *con i quali può essere scambiata l'informazione* ossia il messaggio che si vuole trasmettere. Questa è la definizione più corretta.

Poiché tuttavia "media" in inglese ha anche il significato di mezzi trasmissivi, è possibile cadere in un equivoco: ancora adesso qualcuno pensa che la multimedialità sia la metodologia per cui *una stessa informazione è trasmessa su diversi mezzi* (ossia che equivalga a concepire uno stesso programma per CD Rom,

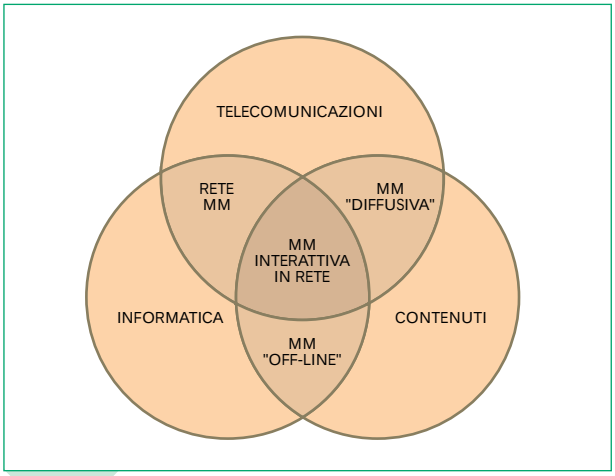


Figura 8 *Focalizzazione sulle aree di convergenza tra le telecomunicazioni, i contenuti e l'informatica.*

per cavo o per satelliti). Anche questa definizione può intendersi come una forma di multimedialità.

Quali sono gli *elementi peculiari* che hanno reso possibile la *rivoluzione multimediale* e l'avvento della Società dell'Informazione?

Anzitutto la *numerizzazione di tutti i tipi di segnali* e in particolare di quelli televisivi; in secondo luogo la *compressione* degli stessi; ancora la *trasmissione* dell'informazione numerica per *pacchetti* e infine, nel caso di Internet, la nascita di un *protocollo universale* di comunicazione (TCP/IP). Questo protocollo rappresenta uno standard "de facto" per la comunicazione tra calcolatori e ha permesso - cosa fino a non molti anni fa ritenuta impossibile - il collegamento tra un computer IBM ed uno Macintosh.

Per comprendere meglio la multimedialità interattiva è opportuno osservare il diagramma di figura 8. Come si vede, si hanno diverse possibili zone di sovrapposizione dei tre cerchi rappresentativi della convergenza: se, ad esempio, combiniamo il "contenuto" con le "telecomunicazioni", ne scaturisce l'attuale sistema di *video diffusivo* ("broadcasting"); se combiniamo le "telecomunicazioni" con l'"informatica" abbiamo le reti multimediali; se combiniamo il "contenuto" con l'"informatica" abbiamo la *multimedialità off-line* (ossia il CD Rom).

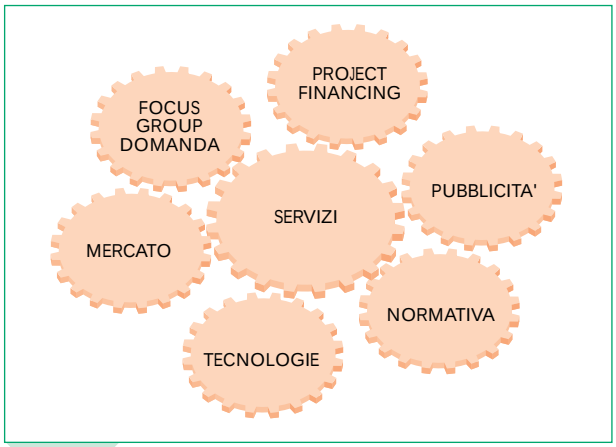


Figura 9 *I servizi come elemento centrale della multimedialità.*

Tutte le combinazioni citate sono forme di multimedialità relative alla prima definizione allargata data in precedenza. A mio parere tuttavia, l'elemento veramente caratterizzante della Società dell'Informazione è la zona centrale comune ai tre cerchi ossia la *multimedialità interattiva di tipo personalizzato* ("on line").

Quando il Vice Presidente americano Al Gore introdusse il concetto delle autostrade elettroniche *(electronic superhighway)* aveva un'idea ben precisa: che da New York un singolo utente potesse interrogare una banca dati di San Francisco per ricevere, ad esempio, un film d'autore. In questa ottica il sistema di trasmissione a lunga distanza trasporterebbe per un singolo utente un'informazione a larga banda per un tempo notevole e quindi sarebbe auspicabile un costo estremamente basso per il "bit" trasmesso. Purtroppo, mentre il costo del "bit" elaborato è crollato, il costo di quello trasmesso è ancora alto.

L'elemento cardine della multimedialità non è tuttavia né la tecnologia né la normativa né la pubblicità né il modo di finanziarli, ma il *vero ed unico "driver" è rappresentato dai contenuti e da un loro uso personalizzato* derivante dalla creazione di nuovi servizi (figura 9).

Opportunità e rischi della Società dell'Informazione

L'*opportunità* maggiore che si presenta nel futuro riguarda certamente l'*elaborazione della conoscenza*. In particolare questa sfida ha per la cultura d'impresa importanti conseguenze che in Italia non sono ancora state comprese completamente.

È quindi necessaria una revisione completa dei processi d'impresa in quanto è controproducente applicare nuove metodologie ai processi così come in passato erano stati concepiti: occorre perciò fare uno sforzo per un completo "business reengineering".

Altre opportunità sono: *l'interazione tra sistemi e individui, la trasparenza, la democrazia dell'informazione.*

Passiamo ora ad esaminare i rischi. Va anzitutto sottolineato che purtroppo l'elenco dei rischi è molto più lungo di quello delle opportunità. In primo luogo, a livello individuale, si presenta un qualche pericolo di una *virtualizzazione della realtà* e della verità per un'eccessiva dipendenza dai computer. Per fare un esempio pratico: molti individui si fidano ciecamente dei dati elaborati da un computer senza essere ulteriormente critici nel fornire ad esso un accurato *modello* del problema esaminato.

Un altro punto importante è il *pericolo di naufragare in un oceano di troppe informazioni*, con il risultato di un isolamento culturale e sociale. Potremmo ricordarci che "tutta l'informazione del mondo" o "nessuna informazione" sono esattamente la stessa cosa!

Passiamo ora ai pericoli a livello sociale ed economico: il primo e più grave è *l'aumento di disoccupazione*, forse temporaneo, ma che in Europa (più che negli Stati Uniti) potrebbe anche essere di lunga durata.

Un'altra sicura conseguenza della transizione sarà l'epidemia di imprese non sufficientemente flessibili e rapide a convertirsi. Questo fenomeno è tipicamente dell'Europa dove l'incremento di produttività del sistema, conseguente all'informatizzazione, è stato nettamente inferiore alle attese.

Può essere ricordata infine un'ultima serie di rischi a livello istituzionale e politico. In particolare la *possibilità di controllo degli individui; la vulnerabilità di sistemi così complessi* (e quindi il pericolo di enormi interruzioni, "black out", dagli effetti devastanti); *la tentazione di un ricorso continuo a referendum* che, in effetti, potrebbero essere facilmente praticati per via telematica; e via di seguito.

Conclusioni

E finalmente giungo alle conclusioni: ho riesumato un "cartoon" (figura 10) di circa venti anni fa in cui Linus dice alla sorellina: "Se guardiamo continuamente la TV non dovremo imparare a leggere, se usiamo il wordprocessor e il calcolatore non dovremo imparare a fare i conti. Abbastanza presto non dovremo imparare più nulla". A questo punto la sorellina risponde: "questo sarà il momento buono per me!".

Questa frase caratterizza uno dei maggiori pericoli temuto dai pessimisti. Come la muscolatura umana negli ultimi cento anni si è notevolmente indebolita per la sua minor necessità di impiego e per l'eliminazione della fatica fisica, così potrebbe accadere per il cervello umano ...

Ricordo ancora una citazione che non mi stanco di riprendere: nel 1934 Thomas Elliot, autore dalle idee penetranti e di intelligenza lucidissima, scriveva in un suo poema:

"Dov'è la saggezza che abbiamo perso nella conoscenza?

Dov'è la conoscenza che abbiamo perso nell'informazione?".

Figura 10 *Linus, da "Le frontiere dell'informatica" di F. Filippazzi e G. Occhini.*

In effetti questi versi sintetizzano le tappe fondamentali dell'umanità: in antico la saggezza; si è poi passati alla conoscenza e in tempi moderni dalla conoscenza all'informazione.

Sorge spontanea allora la seguente riflessione: se pensiamo che il peso delle componenti negative possa essere determinante, perché ci occupiamo della multimedialità?

La risposta a questa domanda è estremamente precisa: *le grandi rivoluzioni della storia*, anche se dense di pericoli, *o si cavalcano o se ne rimane esclusi* con conseguenze ancora peggiori dei pericoli paventati. Occorre quindi essere protagonisti di queste rivoluzioni perché se non si è protagonisti si è vittime.

Poiché come abbiamo già detto tutta l'informazione del mondo o nessuna informazione sono la stessa cosa, possiamo porci una domanda, forse un po' retorica, alla quale non sono in grado di dare una risposta. *"È possibile che nel mondo moderno si riesca a trasformare una parte dell'informazione in conoscenza?"*.

È una riflessione che lascio a voi da approfondire.

Grazie dell'attenzione.