DISSERTATION

Perception Oriented, Delay-Controlled Echo Cancellation in IP based Telephone Networks

Ausgeführt zum Zwecke der Erlangung des akademischen Grades eines
Doktors der technischen Wissenschaften
unter Anleitung von

Univ.-Prof. Dipl.-Ing. Dr. Gottfried Magerl
Institut für Elektrische Mess- und Schaltungstechnik

und

Univ.-Prof. Dr.-Ing. Harmen R. van As
Institut für Breitbandkommunikation

eingereicht an der Technischen Universität Wien
Fakultät für Elektrotechnik und Informationstechnik

von

Dipl.-Ing. Wolfgang Brandstätter
Matr.-Nr. 9425817
Wieshäusl 13
A-4625 Offenhausen

Wien, im April 2004

Abstract

Modern speech communication reveals a strong trend from the well proven circuit towards upcoming packet switched networks. Beside substantial cost savings and exploitation of synergies through convergent networks, packet switching establishes a basis for new, value-adding services, which arises from the opportunity to combine diverse types of media such as video, text, pictures, and audio. The underlying technology of packet switched telephone networks is inherited from the Internet, which has been originally designed for the reliable transmission of time-insensitive data. Therefore, real-time applications experience significant and unpredictable end-to-end delays, which directly impair the perceived voice quality. Another disturbing factor is the unavoidable echoes, which degrade the user's satisfaction with increasing delays. Hence, echo cancellers are facing stringent requirements, which are only partly met especially under double talk conditions.

The delay-controlled enhancement of the conventional principle of echo cancellation improves the perceived voice quality by lessening the extent of influence on the voice signals to be transferred with decreasing echo round-trip times. The signal at the ingress of the echo cancellers is made up of the wanted signals of the local participant and the undesired echo components originated by the distant talker. The control algorithm of the newly developed idea is based on two continually observed parameters. On the one side, the delay of the voice signals from the talker's mouth to the talker's ear determines the overall required echo attenuation according to a predetermined characteristic. On the other side, the measured reduction of the signal levels along the echo path results in—in combination with the aforementioned echo attenuation—the residual echo attenuation needed for the echo canceller. The determination of the echo round-trip delay represents the most demanding challenge within the scope of this work. In this context, the discussed methods of resolution mainly take advantage of the timestamps placed in the transferred voice packets.

The results obtained from a listening test confirm that there is room for improvement in the delay-controlled approach. The modeling of two analog subscriber lines, which are connected by an IP based telephone network, acts as basis for the simulation of various network conditions as well as for the optimization of diverse parameters of the echo canceller. Different aspects of the binaural voice samples, which have been created in this manner, have been evaluated by a panel of subjects in a listening test. The results conclude considerable improvements in terms of voice quality compared to the conventional approach. As the delay decreases the voice quality, under certain conditions, increases to a level equivalent to that of an a priori echo-free telephone connection.

Kurzfassung

Die moderne Sprachkommunikation zeigt einen starken Trend, der von der bewährten Leitungsvermittlung ausgeht und in Richtung aufkommender Paketvermittlung weist. Neben den erheblichen Kosteneinsparungen und der Nutzung von Synergieeffekten durch ein konvergentes Netz, ermöglicht die Paketvermittlung auch die Einführung neuer Mehrwertdienste, die durch die Kombination verschiedener Medientypen wie Video, Text, Bilder und Audio entstehen. Die zugrunde liegende Technologie paketvermittelnder Telefonnetze basiert auf dem Internet, das grundsätzlich für die zuverlässige Übertragung zeitunempfindlicher Daten entworfen wurde. Daher erfahren Echtzeitanwendungen erhebliche und auch nicht vorhersagbare Ende-zu-Ende Verzögerungen, die direkt die subjektiv wahrgenommene Sprachqualität beeinträchtigen. Ein weiterer Störfaktor stellen die mit steigenden Laufzeiten zunehmend als störend empfundenen und auch unvermeidbaren Echos dar. Aus diesem Grund stellt man hohe Anforderungen an die verwendeten Echokompensatoren, denen vor allem unter Gegensprechen nur teilweise entsprochen wird.

Die laufzeitgesteuerte Erweiterung des herkömmlichen Prinzips der Echokompensation verbessert die durch den Teilnehmer wahrgenommene Sprachqualität, indem es mit kleiner werdenden Echolaufzeiten weniger stark in das zur Übertragung anstehende Signal eingreift. Das Eingangssignal des Echokompensators setzt sich aus dem Nutzsignal des lokalen Teilnehmers und den unerwünschten Echos des fernen Sprechers zusammen. Die Steuerung des neu entwickelten Konzeptes basiert auf zwei kontinuierlich erfassten Parametern. Einerseits bestimmt die Echolaufzeit vom Mund bis zum Ohr des Sprechers nach einer vorgegebenen Kennlinie die notwendige Gesamtdämpfung des Echos. Andererseits ergeben die gemessenen Dämpfungswerte entlang des Echopfades gemeinsam mit der zuvor ermittelten relativen Pegelreduktion die benötigte Restechobedämpfung des Echokompensators. Die Bestimmung der Echoumlaufverzögerung stellt die größte Herausforderung im Rahmen der vorliegenden Arbeit dar. Die diskutierten Lösungsansätze machen sich in diesem Zusammenhang großteils die in den übertragenen Sprachpaketen enthaltenen Zeitstempeln zu nutze.

Die Ergebnisse eines Hörversuches bestätigen das Verbesserungspotenzial des laufzeitgesteuerten Ansatzes. Die Modellierung zweier analoger Teilnehmeranschlüsse, die durch ein IP basierendes Telefonnetz miteinander verbunden sind, diente sowohl als Grundlage für die Simulation verschiedener Netzzustände als auch für die Optimierung diverser Parameter des neuen Echokompensators. Die derart erzeugten, binauralen Sprachproben wurden von Versuchspersonen nach verschiedenen Gesichtspunkten in einem Hörversuch beurteilt. Die Ergebnisse zeigen beachtliche Qualitätsgewinne gegenüber dem herkömmlichen Ansatz, die mit abnehmenden Laufzeiten zunehmen und unter bestimmten Bedingungen sogar den Qualitätsstandard einer von vornherein echofreien Telefonverbindung erreichen.

Acknowledgement

I would like to express my gratitude to the many people who have played an outstanding role for the success of this thesis and to those who have directly had a hand in the research work.

My sincere thanks go to my thesis advisor and mentor, Univ.-Prof. Dr.techn. Dipl.-Ing. Gottfried Magerl, for his warm personality, continual encouragement, support, and perseverance in guiding me through the entire research process and for his helpful hints in writing the dissertation—especially for advising me to draft the thesis in English. I am indebted to HR Dipl.-Ing. Günther Fiedler for his multi-functional assistance, ranging from the origination and organization of the research project to giving me the opportunity to present the results of my work at several conferences and workshops. I am grateful to the staff and to my colleagues at the Institute of Electrical Measurement and Circuit Design for their friendship over the past few years and their participation in the long-lasting and exhausting listening test.

Special thanks go to Univ.-Prof. Dr.-Ing. Harmen R. van As for the taking the time to pass on his expert opinions and for giving constructive criticisms on my written work.

Special acknowledgement is also extended to Telekom Austria as well as AOSA Telekom for their financial funding of the research project, for giving me the chance to work on an emerging and utmost dynamic topic and for exercising patience while the theoretical approach with the already mentioned listening test was being confirmed. Beside monitoring the project and acting as subjects in the experiment, they also provided valuable feedback and suggestions to my research work and delivered insight into the telecommunications industry in weekly meetings.

I also had the fortunate opportunity to get known to Dipl.-Ing. Frank Kettler from Head Acoustics, who gave essential inputs to the novel approach of the delay-controlled concept. Moreover, he has been supportive in designing and giving me useful advices for the auditive investigation. His contribution is greatly appreciated.

It would be beyond the scope of this acknowledgement to exhaust all the friends, who encouraged and supported me throughout the research and writing process or helped in the conduction of the listening test. They have always been a source of inspiration, help and delight.

Last but not least, my heartfelt thanks go to my family for their continual understanding as well as their unconditional and moral support—they have been providing me a place of ease and comfort and a source of energy my whole life.

Contents

Chapter 1

Introduction

> The origins of the word echo: „In Greek mythology, a mountain nymph, or oread. Ovid's Metamorphoses relates that Echo offended the goddess Hera by keeping her in conversation, thus preventing her from spying on one of Zeus' amours. To punish Echo, Hera deprived her of speech, except for the ability to repeat the last words of another. Echo's hopeless love for Narcissus, who fell in love with his own image, made her fade away until all that was left of her was her voice."[1]

1.1 Paradigm Shift

Traditional telephony networks have been conceived to offer an optimal service for time-sensitive voice applications requiring low delay, low jitter, and low but constant bandwidth. Such requirements are realized by arranging a connection, which is exclusively dedicated to the participants at both ends of the line over the whole duration of the call. The excellent quality in terms of reliability, availability, and, above all, clarity comes, adversely, with high acquisition and operation expenditures.

The Internet, on the other hand, lessens these drawbacks significantly, due to the ever emerging spread of the underlying Internet protocol (IP) technology, which drives the network costs down to a fraction of the conventional telephony system's expenses. The general tendency among carriers also points at an integrated IP network for various applications, which are comprised not only of the Internet, but also of voice, video, television, and

---

[1] Encyclopædia Britannica.

combinations of these basic services. These synergies implicate significant economic benefit for network operators. Furthermore, new promising services, which combine the capabilities of the Internet with those of classical voice applications, are realizable. Since the Internet has been originally designed to deliver non-real-time traffic such as e-mail or file transfer in a reliable way, its basic principle consists of guaranteed end-to-end transfer of occasionally bursty traffic without considering the effective delay. For this reason quality-assuring mechanisms have to be added to the Internet, which prioritize the delay-sensitive voice traffic over other data.

Contrary to the expectations, the great potential shown by an IP based telephone network has only been exploited to a little extent so far. Although carrier solutions are already capable of fulfilling the stringent requirements of professional network operators, the great economic slump in the information technology sector has forced them to lower their expenditures. Therefore, the equipment vendor's sales experienced a resultant downturn, as well.

In spite of this, the common consent within the telecommunication industry leans towards an IP based network, which shall be deployed in the backbone as a first move in order to utilize the two-wire copper line as long as possible. New narrowband services are brought in easily in such an environment. Only greenfield operators are in the position to build up an all-IP network, since they do not have to exploit existing acquisitions. The speed of the technology shift towards an IP based telephone network will strongly depend on the future development of the business situation.

1.2 Motivation

The nature of voice transfer over IP networks consists in collecting speech samples and forming compressed packets, which are routed along several network nodes to the receiver where they are converted back into a continuous voice stream. On this account Internet telephony is always faced with a lower limit of end-to-end delay. Results of test events involving carrier-grade equipment pointed out minimal time spans from one gateway to the other of at least 50 ms [155]. These values are well beyond the required quality degrading threshold [70], but in the case of long distance calls and/or connections over satellites or other radio communication systems as well as delay-adding units such as IP phones, the resulting one-way delay may be above 100 ms, which hampers the turn-taking in a dialogue. Moreover, if there are no prioritization schemes installed in the network the overall delay varies and may become unpredictably high due to network congestion.

Besides the non-negligible end-to-end delay, the talkers are also faced with unavoidable signal reflections of their own voice arising at the interfaces where the two-wire customer loop inserts the voice signals into an IP network. In fact, this conversion point is placed in the user's local switch, where the analog signals of the two-wire line are separated for both transmission directions, before they are grouped in, and finally sent as, IP packets by the gateway. Such echoes are unavoidable in the described and widely deployed scenario. The grade of the user's annoyance significantly depends on the mouth-to-ear delay as well as the difference between the original signal level at the talker's mouth and the reflected level at the talker's ear. As IP telephony, in principle, comes with significant signal retardations, echo problems are present in such communication systems in most cases. Therefore, commercially available gateways are always equipped with a so-called echo canceller, which is, for example, also encountered in international switches and has turned out to be indispensable in mobile switching centers (MSC). The results of various tests and trials, as already indicated above, have emphasized that echo control devices implemented in present gateways affect the transferred voice signals considerably—especially when both parties are talking at the same time. The latter condition of a communication is referred to as double talk.

Due to these factors, a new approach on echo cancellation is presented within this thesis, which reduces the impairments of perceived voice quality in general, and particularly in double talk situations.

1.3 Task Description

The echoes are less disturbing as the mouth-to-ear delays decrease. Traditional implementations of echo control devices handle echo signals independently of the round-trip time. Hence, even in the case of low delay values of about 10 ms to 20 ms, where echoes are desired for a comfortable talking situation, they are trying to compensate and, under some circumstances, also degrade the wanted signal induced by the participant at the local end.

Therefore, the new concept on echo cancellation enhances the basic principle of standard designs by considering the round-trip delay of the echo. Since the non-linear component of the canceller, which is discussed in depth in Section 2, is responsible for the distortions in the opponent's voice, the newly introduced time parameter directly alters its control behavior. This is accomplished by determining the required overall echo attenuation from the measured round-trip delay and, based on this value, the resulting attenuation needed for the non-linear function is derived. For this purpose the current reduction of the echo level introduced by the talker's telephone, by the two-to-four wire conversion unit, and by the adaptive filter, which is also found as a basic part in every echo canceller, has to be monitored continually.

A very crucial objective in this context is the detailed discussion and comparison of the possibilities for determining the required mouth-to-ear delay with adequate accuracy along the echo path. Moreover, the continuous measurement of the attenuation of the reflected signal turns out to be another critical issue.

The delay-controlled concept is verified by a listening test. First, the telephone system, consisting of a circuit switched network with an IP backbone as well as two analog telephones and echo cancellers, is modeled. Secondly, certain network conditions and different settings of the echo canceller are varied and simulated based on the model. Thirdly, the output of the simulations is taken as input for the binaural recordings by an artificial head measurement system. Finally, a panel of subjects rates the different aspects of the prerecorded voice samples.

1.4 Outline

Chapter 2 starts with the basic principle of an IP based telecommunication system, which is characterized in terms of the overall architecture, the functional components and the protocols utilized on the corresponding interfaces between the entities. After a general introduction in voice quality, the particularities of the delay in such networks are discussed; furthermore, echoes and means of echo cancellation are presented. The main focus centers around the ascertainment of the voice quality and the identified problems with echo control devices.

Chapter 3 presents the novel concept on echo cancellation consisting of a perception oriented and delay dependent control of the non-linear component. The main part deals with methodologies for measuring the echo delay, which are either based on the examination of the transferred signal or they rely on the timestamps, which are delivered with every voice packet.

Chapter 4 covers the realization of the listening test carried out to confirm the new approach and, moreover, optimize some parameters of the echo canceller. The assessed voice samples are created by modeling and simulating the needed IP network conditions before the evaluation procedure is carried out.

Chapter 5 addresses the analysis of the individual votes and shows the resulting figures. Results gathered under double talk conditions form the principal part, while the optimization of the double talk detector and the appraisement of echo reference samples rounds out this chapter.

Chapter 6 gives a short outlook on the realization of the discussed idea and concludes the thesis with a short summary.

Chapter 2

State of the Art in Echo Cancellation in IP based Telephone Networks

2.1 Voice Quality in IP based Telephone Networks

> Packet switching is defined as “the process of routing and transferring data by means of addressed packets so that a channel is occupied during the transmission of the packet only, and upon completion of the transmission the channel is made available for the transfer of other traffic.”[2]

2.1.1 Introduction

Packet switching is the underlying technology for computer networks and has become an attractive option for transmitting human speech on a real-time conversational basis. Transmitting voice over data networks promises opportunities for organizations to reduce costs and enables new applications. The proliferation of the Internet[3] in the last years has been responsible for a strong interest in the option of carrying real-time voice traffic over

---

[2] Glossary of Telecommunication Terms-Federal Standard 1037C.

[3] A worldwide interconnection of individual networks operated by government, industry, academia, and private parties. It has been built upon the Defense Advanced Research Projects Agency (DARPA) networks, the National Science Foundation Network (NSFnet), and other regional and national networks.

the Internet. The Internet uses packet-switching and is based on the connection-oriented transmission control protocol and IP (i.e., TCP/IP) protocol suite. Since it was not originally designed for real-time communications, carrying voice over the Internet introduces a number of challenges and technical issues, which have to be worked out before successful deployment of telephony over the Internet may come true. Some of these technical challenges comprise the lack of guarantee in terms of bandwidth, packet loss, delay, and packet delay variation which may influence significantly the quality of voice over the Internet.

On the other hand, the existing telephone system relies on circuit switching and is called public switched telephone network (PSTN). A connection (circuit) has to be set up between two end-points before the start of communication. The channel is exclusively dedicated to the participants at both ends of the line during the whole call, independent from the amount of traffic transferred over the connection. Thus telecommunication networks offer a guaranteed quality of service (QoS) to customers. The main advantages of this analogue and so-called plain old telephone service (POTS) are summarized as follows [9]:

- Worldwide availability.
- Very high reliability.
- Excellent voice quality.
- Ease of use.

The major drawbacks are the high acquisition and operating costs. When building up a telecommunication network, one has to consider the costs for the separate wiring, high acquisition costs for the telephones and switching components, high expenditure for the staff as well as the operating costs due to an external service contract and much maintenance work.

Drivers

The main forces for the deployment of IP based telephone systems are the opportunity to realize new, value adding services and to save costs. Telephony over the Internet enables the creation of a number of new *services* and their integration which would not be possible using traditional circuit-switched telephone networks. The combination of voice services with data applications like text, fax, audio, pictures, video, and others over the same medium creates this new set of functions offered to the user. A few of the emerging applications enabled by this new technology are [56]:

- Video telephony.
- Improved voice quality: IP telephony can support higher grades of sound than the PSTN by deploying broadband coding with a frequency range up to 7 kHz. On the other hand the term toll-quality provided by the PSTN defines "the voice quality resulting from the use of a nominal 4 kHz telephone channel".[4]
- Unified messaging: The user receives different kinds of messages (e-mail, phone, voice mail, and fax) at one location from where they are accessible.

---

[4] Glossary of Telecommunication Terms-Federal Standard 1037C

- Web-based call centers: They allow subscribers browsing the Internet to initiate an IP based call from an organization's web site to its call center. The Internet surfer does not need to stop browsing; instead, the call enables just an extension of the user's web activities.
- Real-time billing: The IP telephony user can access the gateway for billing information in real time.
- A virtual second line: With IP telephony, subscribers utilize the same line for voice calls and for Internet surfing.
- Remote tele-working and distance learning applications.
- Enhanced teleconferencing using shared applications and whiteboard.

The second driver is *cost saving*, which is possible because of the following reasons:

- The operation and maintenance of only one unified communication network, which is referred to as convergent network or next generation network (NGN), is less costly than in the case of separated networks, e.g. the transport of intra-company voice communications between remote sites over the data network of that enterprise.
- Unified messaging with IP telephony reduces system and network complexity.
- The available bandwidth is utilized more efficiently than in circuit switched networks. No communication link is dedicated to voice calls as within the PSTN. All calls share the same network resources. Such sharing significantly brings down the cost of a phone call. Additionally, the emergence of reduced bit-rate voice compression enhances the carrying capacity of a network by nearly ten times without the investment of additional resources.
- IP telephony offers very good scalability through the connection of additional IP phones to the IP network.
- The acquisition and maintenance of packet switches is less costly than that of the switches applied in the PSTN.

Beside the new services and the business drivers the increased deregulation as well as the steady growth in data network investment complete the drivers for IP based telephony.

Definitions

Communication networks are categorized into two basic types: Circuit-switched (sometimes called connection-oriented) and packet-switched (which are operated connectionless or connection-oriented).

In packet-switched networks, data to be transferred across a network is segmented into small blocks called *packets* (also called datagrams or protocol data units) that are multiplexed onto high-capacity connections. A packet is generally defined as "an information unit identified by a label at layer 3 of the open system interconnection (OSI) reference model" [91]. More precisely a packet is understood as: "In data communication, a sequence of binary digits including data and control signals that is transmitted and switched as a composite whole."[5] The control portion of the packet contains information that enables

---

[5] Glossary of Telecommunication Terms-Federal Standard 1037C

the network hardware to know whereto send it in order to reach the specified destination. In frame relay (FR), the basic transfer unit is the data link layer *frame*. In asynchronous transfer mode (ATM) networks, this basic unit is called the data link layer *cell*. FR and ATM are low-level network transport mechanisms based on packet-switching, which come with some circuit-switching principles like call setup mechanism. Therefore, they are called connection-oriented, which generally means that communication proceeds through three well-defined phases: connection establishment, data transfer, connection release. TCP, for example, is a connection-oriented protocol. On the contrary, IP provides a connectionless means of transport, as there is no call setup; each packet finds its own way across the network independently of the previous one.

The often used term *voice over IP* (VoIP) defines the transport of voice samples over an IP network and, thus, refers to the transport technology. Both Internet telephony and IP based telephony rely on VoIP networks. The term *Internet telephony* points out the utilization of the public Internet, which does not guarantee a certain level of QoS. The Internet treats every packet in the same way and, thus, it serves over a broad range, from no throughput due to totally congested router buffers up to very fast packet transfers. This general behavior of the Internet is called best effort. On the other hand the term *IP based telephony* (or IP telephony) gives emphasis to the service offered to the user. IP telephony draws on a high-quality private network or on a public IP network with guaranteed quality bounds; both networks are, consequently, named in this thesis IP based telephone network.

Alternative Transport Mechanisms

As already mentioned, FR and ATM also belong to the family of packet switching technologies. FR is a layer two protocol and has been built on the X.25 packet-switched technology. It is commonly understood as virtual leased line between two end-points [18]. FR provides a low overhead compared to IP, and it is capable to transfer frames of varying sizes rapidly, whereby the bandwidth allocation occurs dynamically. Furthermore, it has been designed for the data transport in wide area networks (WANs), e.g. the transmission of IP packets over FR networks [127]. FR comes with implemented QoS mechanisms and is quite cheap to deploy. The major drawbacks are that FR only works with data rates below 2 Mbit/s and that different frame lengths may introduce high delays. These are the reasons why FR has been increasingly replaced by TCP/IP and ATM networks, and why it will completely disappear from the market in the future.

ATM uses a fixed-size cell length of 53 bytes (48 bytes payload and 5 bytes header) [18]. It provides an efficient means of traffic management by offering extensive QoS mechanisms and service classes. ATM is able to transport voice with high quality and without any delay variations. Furthermore, delay can be kept quite low in comparison to VoIP networks. ATM has not succeeded in local area networks (LANs), because of the less costly Ethernet technology. Carriers deploy ATM in WANs in order to transport voice, data, video, and digital subscriber line (DSL) traffic over optical transport mechanisms. An advantage of ATM is its scalability, i.e. when upgrading the ATM equipment much more traffic can be carried over the same optical fiber. Recently, ATM has even been partly substituted in WANs by the cheap TCP/IP technology. But ATM will play an important role in the future, as there has been made large investment to deploy it in backbones and it is able to transport voice with excellent QoS. Nowadays the most amount of TCP/IP traffic in WANs is carried over ATM networks. There are trends to carry TCP/IP directly over optical transport technologies because of saving costs and the large overhead of IP over ATM.

Internet telephony—as already stated—offers no guarantees on the achievable QoS level. In spite of this, IP has become the de facto standard connectionless packet network layer protocol for both, LANs and WANs. Nowadays it is already feasible to attain acceptable service quality in corporate networks with modern network technologies like broadband backbones and voice packet priorization. However, one has to consider that data networks only have a low level of availability and reliability compared to the PSTN.

2.1.2 Architecture and Protocols

The work within several standardizing organizations has been on the specification of an integrated and convergent network architecture (i.e., the NGN scheme) and on the corresponding protocols on the interfaces. The Multiservice Switching Forum (MSF) [160], the International Softswitch Consortium (ISC) [158], and the European Telecommunications Standard Institute (ETSI) project Telecommunication and IP Harmonization over Networks (TIPHON) [161] have mainly focused on such architectures and they have used specifications of the Internet Engineering Task Force (IETF) and the International Telecommunication Union-Telecommunication Standardization Bureau (ITU-T) as underlying documents for their outputs. When incumbent carriers decide to evolve the circuit-switched network into a packet-switched one, they often think, as a first step, of the so-called, virtual trunking scenario (see Figure 2.1). In this context the IP network provides “virtual” trunks in the backbone between the switches and the behavior of the system is transparent from the PSTN/ISDN (Integrated Services Digital Networks) point of view, i.e. the transition to and the introduction of IP technology takes place unnoticed by the user in most cases. All services, available to the user before the migration, should remain and perform as before. The illustration in Figure 2.1 depicts an architecture composed of PSTN/ISDN facing the user line and the IP network.

Each architectural function shown in Figure 2.1 is either realized as separate physical device or some of them are summarized in one unit. For example, the first concept of gateways included both the media gateway and the media gateway controller function. The user in Figure 2.1 has access to the network via an analog phone or via a fully-digital connection using an ISDN appliance. The components and implemented functions in Figure 2.1 are described as follows:

The call controller[6] (also called media gateway controller, softswitch, call agent, or gatekeeper[7]) plays a key role in NGN. Its wide variety of functions may comprise connection control, protocol translations, routing, gateway management, call control, bandwidth management, signaling, provisioning, security, and call detail record generation. One essential task is the control of the media gateways via the media gateway control protocol

---

[6] The term call controller has been introduced for an unbiased explanation of the NGN example, as the term call agent is preferred by the MSF, while the term softswitch has been introduced by the ISC.

[7] A gatekeeper accomplishes a subset of the call controller functions in an H.323 environment.

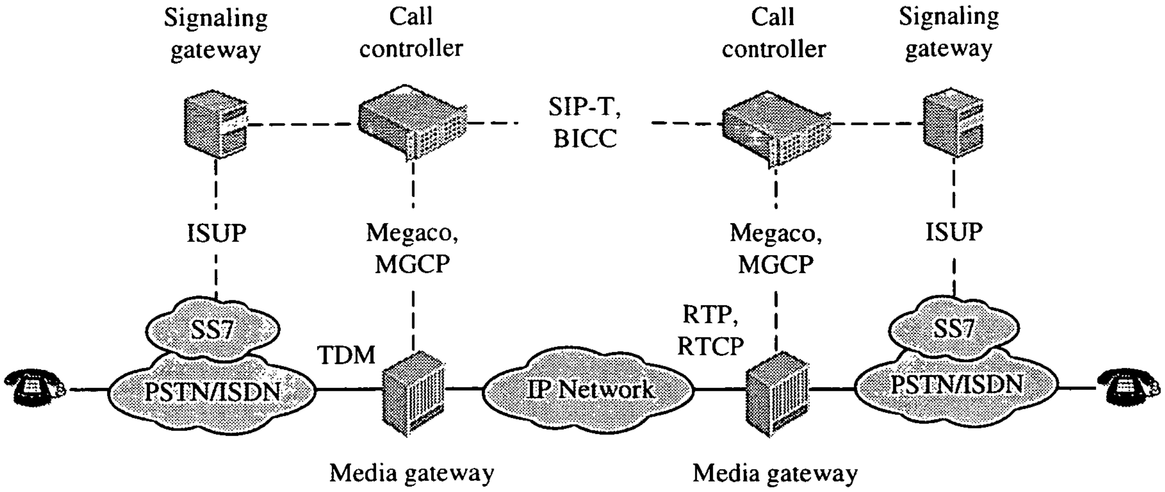

Figure 2.1 – Example of two PSTNs/ISDNs connected via an IP network. The components of the NGN architecture communicate via protocols applied at their interfaces.

(MGCP) [140] or via the emerging Megaco/H.248 protocol [141] [90]. The translation of Signaling System No. 7 (SS7) messages into appropriate IP signaling messages and vice versa is incorporated in the call controller or fulfilled by a separate component—the *signaling gateway*. The bearer independent call control (BICC) protocol [111] or the session initiation protocol (SIP) [143] for telephones (SIP-T) [144] handle the communication between two call control entities, whereas the BICC protocol, beside the IP, may also be used with other underlying network technologies like ATM. There are also functions not explicitly shown in Figure 2.1: The call controller may access media server (for specialized media resources like interactive voice response, conferencing, fax, announcements, and speech recognition systems), application server (for the execution and management of enhanced services or for creating and deploying services via application programming interfaces), or services made available by the Intelligent Network (IN) and accessed via the signaling gateway or in a direct way. The signaling of IP clients and IP phones[8] via the H.323 [89] protocol family or via SIP may also be performed by the call controller. Those end-points of the communication channel are directly connected to the IP network. In addition to that there are interfaces to the management unit using appropriate protocols.

Beside the signaling gateway there exists another fundamental gateway element. The *media gateway* (more specifically called access, residential or trunk gateway) represents the interface between circuit-switched resources (lines, trunks) and the packet network (IP, ATM). Its main functions are voice compression, compensation of IP packet delay variation and IP packet loss, fax relay, echo cancellation, and digit detection, which are, among others, explained in detail in the next Section 2.1.3. The PSTN/ISDN is connected via time

---

[8] At the beginning IP clients were the only available type of device for the end user. They are running as software on Multimedia PCs, which are computers equipped with sound card, microphone and loudspeaker or headset. The first software provider has been the Israeli company Vocaltec [158] in 1995 [9]. Later on, the first hardware devices appeared in 1998 as so-called IP phones. Compared to the telephony software they have drawn from significantly lower delay values, but the costs compared to IP clients are also much higher.

division multiplexing (TDM) links to the gateway.

On the packet-switched side of the gateway the voice samples are accumulated and grouped into *real-time transport protocol* (RTP) [129] packets for the transfer over the IP network. The RTP performs the end-to-end transmission of time-sensitive traffic like interactive voice or video applications over IP networks. It is typically used on top of the connection-less user datagram protocol (UDP) because the TCP transmission scheme is not adapted for data that needs to be carried with very low delay. The IP, UDP, and RTP header of the corresponding protocol stack have a length of 20, 8, and 12 bytes, respectively, i.e. the IP network adds a total overhead of 40 bytes to the real-time packets. The RTP allows compensating for the packet delay variation and corrects the desequencing introduced by IP networks, by using the 4 bytes timestamp and the 2 bytes sequence number in the RTP header, respectively. The RTP control protocol (RTCP) [130] is often used with RTP, which allows the conveyance of some feedback on the quality of the transmission (the amount of delay jitter[9], the average packet loss, etc.) and it is also capable to carry some information on the identity of the participants. In the majority of cases RTCP is also packed into UDP datagrams.

2.1.3 Functions

The voice from the talker's mouth has to traverse several devices executing various functions before it finally reaches the listener's ear. Every analog or ISDN phone connected to an IP network via a media gateway as well as IP phones or IP clients perform—more or less—the key functions illustrated in Figure 2.2.

The echo control device is only needed in the case of an analog phone, where a two- to four-wire conversion unit—not shown in the figure above—in the local switch introduces some amount of echo, or when acoustic coupling effects between the loudspeaker and microphone of the phone are occurring. In such cases some amount of the signals on the incoming path from the remote talker is transferred into the sending path of the local party. That's why an echo control device is deployed in order to reduce or eliminate the echo originated by the distant talker's voice. A detailed discussion on echo and echo cancellation is given in Section 2.3.

The voice path starts with the acoustic to electric conversion in the microphone. In the case of a hands-free or mobile environment the background noise is reduced by a so-called noise reduction unit. After that, in the case of an analogue subscriber the previously mentioned two- to four-wire coupling entity in the switch is passed, before the analog voice signals are digitized. They may further be modified by a succeeding echo control unit.

Automatic Gain Control

In a next step, the level of the speech signal is adjusted by the automatic gain control (AGC) unit (also known as active level control [58] entity) to meet the required speech levels in the network [81]. Speech level may have a strong influence on the perceived

---

[9] Jitter is "an undesirable random signal variation with respect to time", from Babylon Digital Television Glossary.

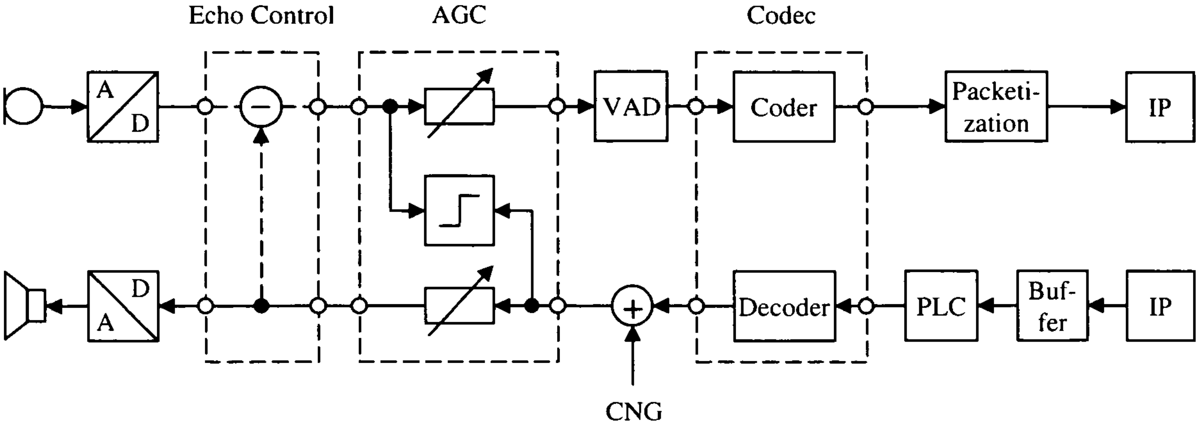

Figure 2.2 – Key functions of an IP based telephony system.

voice quality.

ITU-T Recommendation P.56 [93] specifies a measurement method for a widely used measure, the active speech level (ASL). Within the core PSTN ASLs are expected to be around -18 dB_{m0}[10] and hence the telephone network is optimized for this level. The AGC keeps the speech level in the optimal range; otherwise the AGC would introduce voice degradations.

Silence Compression

During a two-point conversation, subscribers talk only 35 percent of the time. Thus silence compression or suppression enables considerable bandwidth savings of about 50 percent in a standard point-to-point call. In decentralized multicast conferences the activity rate of each speaker drops and the savings are even greater. Silence compression consists of three major functions: Voice activity detection, discontinuous transmission, and comfort noise generation.

Voice activity detection (VAD) has to distinguish between talking and silent periods of the user. It should react very quickly, because some syllables or even words at the beginning of an active phase may get lost quickly, or useless silence might be included at the end of a sentence. At the same time, the VAD should not be triggered by background noise. VAD assesses the speech level of the incoming samples and activates the media channel if this level is above a minimum. When the level falls below a certain threshold for some time, the media channel is muted. Therefore, the detector functions like a gate that opens upon the occurrence of significant speech at the ingress. If the VAD module drops all samples until the mean energy of the incoming signals reaches the threshold, the beginning of the active speech period is suppressed. This phenomenon is commonly referred to as *front-end clipping*. For that reason VAD realizations need some look-ahead, i.e. they keep a few milliseconds of the speech samples in memory in order to activate the media channel before an active speech period is finally detected. For most implementations this measure increases the end-to-end delay; but some coders are able to minimize that addi-

---

[10] The unit "dB_{m0}" denotes a power level relative to 1 mW when passing through the reference point.

tional time span. A well implemented VAD comes with a minimal look-ahead and still minimizes voice clipping, and has a configurable hang-over period (e.g., 150 ms). As VAD stops the transmission during silence periods, it also lowers the consumed electrical energy and is, thus, useful for battery-powered devices.

Discontinuous transmission (DTX) denotes the capability of a codec to stop transmission of frames when the VAD detects a silence period. Some advanced codecs like G.723.1 or G.729 do not stop transmission completely, but instead switch to silence mode in which they use much less bandwidth and just send the basic minimum parameters like intensity in order to allow the receiver to recreate background noise.

Comfort noise generation (CNG) is a function implemented on the receive-side which is complementary to the VAD. It aims to regenerate some sort of background noise, when the VAD has decided on a silence period. Otherwise the called party would get the impression that the line has been dropped. CNG provides means to avoid this behavior. In the case of a simple codec that just stops transmission, it would use some random noise with a level deduced from the minimal levels recorded during active speech periods. More advanced codecs such as G.723.1 or G.729 have options to send enough information to allow the decoder to regenerate ambient noise close to the original background noise. The match between the generated noise at the receiver and the "true" background noise at the sender determines the quality of the CNG.

Voice Coding

The term codec (also called coder) is the short form of coder-decoder. The primary function of a codec is to convert analog video and audio signals into digital data and vice versa. They also perform speech analysis at the sender with the intention to compress the voice data (i.e., redundant or less important information is removed) and, to reduce the bandwidth requirement for transmission over digital networks in this way. Compression is a balancing act between voice quality, computation power, delay, and network bandwidth requirements. Compressing voice signals is computation intensive. The greater the bandwidth reduction the higher is the computational cost of the codec for a given level of perceived quality. In addition, greater bandwidth savings generally come with a higher end-to-end delay. The network planner must make a trade-off between bandwidth and voice quality. Furthermore, low-bit-rate speech codecs such as G.729 and G.723.1 try to reproduce the subjective sound of the signal rather than the shape of the speech waveform. This means any lost or severely delayed information can have much more noticeable effect on the clarity than with a higher bit-rate speech codec. Digitizing and compressing is performed by the coder, while the reverse process of speech synthesis and converting the received signal back into analog format is carried out by the decoder.[11, 12] The corresponding functions are referred to as coding (or encoding) and decoding. Both are implemented in opposite directions of transmission in the same equipment.

This work assumes the deployment of audio codecs for the use in telecommunication equipment. Most voice codecs take a fixed-size number of linear samples and compress

---

[11] Glossary of Telecommunication Terms-Federal Standard 1037C

[12] Babylon Digital Television Glossary

them into one frame. The length of each portion of voice samples in the time domain is denoted as frame size, which corresponds to 10 to 30 ms of speech for typical frame-oriented coders. The time required for accumulating and compressing the speech frame is defined as coding delay (see Section 2.2.1).

There are two major techniques used in speech analysis, waveform coding and source coding (or vocoding). G.711, for example, is based on a very simple form of waveform coding, the pulse code modulation [82], and is deployed in the PSTN/ISDN all over the world. Detailed explanations on the two major techniques as well as on the commonly used hybrid coding are found in [4] and [15]. Hybrid coding uses aspects of both waveform and source coding, bringing together the benefits of the good quality speech of waveform coders and the low bit rates of source coders. Two widely deployed examples for hybrid coders are the G.723.1 and the G.729. Both are listed in Table 2.4 with their main attributes beside the standard G.711 codec.

One or more frames of the output of the codec are grouped into packets (packetization unit in Figure 2.2) and are then sent to the IP network. The receiver of the packets has to realign the different packets according to their end-to-end delays by the means of the so-called de-jitter buffer (see Section 2.2.1).

Packet Loss Concealment

Before the packets are finally played out, the packet loss may be compensated. This ability of a VoIP device makes a significant difference to its performance. Certain standardized codecs, for example the G.729 or G.723.1 codec, include their own packet loss concealment (PLC, also known as error concealment) methods. However, the use of proprietary PLC may improve the voice quality over these standard methods.

An alternative to PLC is the insertion of silence in the place of lost packets, which results in very annoying clicking effects. It is also possible to add redundancy in order to correct some errors or to recreate lost packets. Techniques available include forward error correction, the duplication of frames across multiple packets, or frame interleaving. The drawbacks of these methods are the increased delay and the higher bit rate. A summary on PLC is given in [56].

2.1.4 Voice Quality Terminology

The meaning of the term *quality* is very broad. In telecommunications it is commonly used whether the service satisfies the user's expectations. The evaluation, however, depends on various criteria related to the party rating the service. Customers assess it on the basis of a personal impression and in comparison to their expectations, while an engineer expresses quality in terms of technical parameters.

The term service in the telecommunications context refers to the ability to exchange information via a telecommunication medium, and it is offered to a customer by a service provider [54]. A service in an IP environment (also named IP-based service) is defined by ITU-T as "a service provided by the service plane to an end user (e.g., a network element or a host[13]), which utilizes the IP transfer capabilities and associated control and management functions, for delivery of the user information specified by the service level agreements" [115]. ITU-T describes parameters, attributes and classes of IP-based services.

There are three notions of QoS within the general QoS model [6], which can be regarded as layered beginning with the assessed QoS on top, then the perceived and finally the intrinsic QoS on the transport level. The assessed QoS is defined as "the collective effect of service performance which determines the degree of satisfaction of a user of the service [64]". This definition has been introduced by ITU-T; ETSI has adopted, more or less, the same description [147]. The assessed QoS depends—among others—on the service price, the responses of the provider to problems and complaints, the reliability and availability of the service as well as the perceived quality. An example of the assessed QoS for voice services is given in Section 2.1.5. The perceived QoS reflects the user's overall opinion on the service during a connection. In this context, the QoS also comprises conversational aspects of the quality (e.g., conversational voice quality in Section 2.1.5). Unlike the ITU-T, IETF has no distinct definition of perceived QoS. IETF focuses on the transport layer and defines the corresponding intrinsic QoS as "a set of service requirements to be met by the network while transporting a flow" [134]. The ITU-T specifies IP performance parameters on the transport level in [116]; values for four of those measures are given in [117] for each of the six QoS classes introduced in the same document. Within both, the IETF and the ITU-T, network performance in packet networks is expressed by the following set of parameters: Available bit rate for the service, experienced delay, variations in the IP packet transfer delay (jitter), packet loss rate, and—in some cases—the bit error rate.

The basic benefit provided by a telecommunication system is the voice service. Subscribers of the PSTN are—as already mentioned in Section 2.1.1—used to a high QoS level. The future success of IP telephony service will strongly depend on how it performs compared to the toll-quality of the PSTN. First, carriers will have to provide basic IP voice telephony service with a guaranteed QoS level. Methodologies for assured QoS performance in IP networks are mentioned in the queuing delay passage of Section 2.2.1. Second, when the technology is well developed and gets wide consumer acceptance, integrated interactive multimedia applications are following.

2.1.5 The Conversational Voice Quality

The main factors affecting the assessed QoS for voice services (also referred to as end-to-end or mouth-to-ear QoS) are divided into speech and non-speech related parameters. The speech part of the assessed QoS generally corresponds to the term *voice quality* (also called speech quality[14]) and, when emphasizing the two-way character of a communication, the term *conversational* voice quality is used. Non-speech parameters are identified before a

---

[13] A host is defined as follows (taken from Glossary of Telecommunication Terms-Federal Standard 1037C): "In packet- and message-switching networks, the collection of hardware and software that makes use of message switching to support, user-to-user, i.e., end-to-end communications, inter-process communications, and distributed data processing ."

[14] The term "voice quality" is preferred within the ITU-T, while "speech quality" is often found in ETSI TIPHON and Speech Transmission Quality [162] specifications.

communication starts and also during a call; they depend on the following factors:

- The particular kind of voice service (e.g., calling card, voice-mail, free-of-charge calls, and call forwarding).
- Price of the service.
- The *availability* in terms of down-time and network busy indications.
- The *reliability* quantified by the percentage of dropped calls, wrong numbers, and calls not completed.
- The call establishment quality mainly characterized by the call setup time; perceived by the user as the responsiveness of the service. The setup time includes the start dial signal delay and *post dial delay* [154].

The conversational voice quality represents the perceived QoS and mainly relies on clarity, echo, and delay; it is categorized as follows [152]:

- Listening quality (also called one-way, assessed or perceived voice quality; or speech clarity[15]): The listening quality evaluates the voice samples originated from the talker's voice at the other side and perceived at the listener's ear. It is dominated by voice distortion and noise; the one-way delay is not considered and talker echoes are not occurring, because of the strict one-way situation for this measure. Speech clarity is expressed by means of some psychological parameters [146]: The intelligibility, the naturalness and the loudness. The *intelligibility* denotes a subjective measure of how much information can be extracted from a conversation. The *naturalness* defines the degree of fidelity of the speaker's voice in order to, for example, identify the speaker or notice nuances such as the speaker's emotional state [55]. The *loudness*, finally, describes the absolute loudness level at the receiver's side. Besides, the listening quality significantly relies on the quality of the background noise transmission.
- Talking quality: The impairments influenced by the talker's own voice are mainly determined by echoes and sidetone distortions (see Section 2.2.1).
- Interactive quality: The quality associated with the alternation of talking and listening (also known as turn-taking). The corresponding parameters in this context are the end-to-end delay (see Section 2.2.2) and noise or speech switching.

Beside the pure listening and the pure talking quality, the combined scenario of talking and listening of both parties at the same time has a major impact on the so-called double talk quality. Under such duplex conditions the echo canceller may degrade the end-to-end voice quality significantly (see Section 2.3.9).

Most of the factors listed above are occurring independently of the underlying network technology (e.g., circuit-switched or IP network). However, IP telephony systems without assured QoS mechanisms come with some special characteristics such as long delays, delay jitter, packet loss, and sometimes limited bandwidth. The sources and consequences of delay and delay jitter are discussed in depth in Section 2.2, while packet loss is treated later

---

[15] A clear distinction between assessed and perceived QoS has been made in this thesis. On the contrary the terms perceived and assessed voice quality have the same meaning and are both referred to the one-way character of the listening quality.

on in this section. To cope with these special attributes there are signal processing components in the mouth-to-ear path that contribute to the end-to-end QoS in a unique way. Among others, these parts are, as already mentioned in Section 2.1.3, the voice coding and silence compression unit, the PLC and/or error-concealment technique as well as the de-jitter buffer implementation. The voice coding scheme in the PSTN (i.e., pulse code modulation or PCM) has a minor impact on clarity through minimally increased signal-to-noise ratio as it is shown by the mean opinion score (MOS) value in Table 2.4. The control algorithm for the de-jitter buffer in IP networks has to find a trade-off between additional delays and dropped packets. The echo control device plays an important role since it has to provide high echo attenuation because of the quite high delays (which makes a certain echo level more annoying, see Section 2.3) in IP networks. Another voice degrading technique used in interfaces between networks with different voice codecs is transcoding or codec tandeming [58]. Each coding and decoding process impairs the resulting voice quality and increases the end-to-end delay. These problems can be avoided by a tandem-free operation, where the systems negotiate a common codec which is used end-to-end.

Packet Loss

In contrast to the PSTN, there are no end-to-end circuits established in IP networks. Therefore, packet loss remains unnoticed and represents a common problem in IP networks. IP packets from many sources are queued for transmission over an outgoing link in a router. These packets are transmitted one after another from the head of the queue (see queuing delay in Section 2.2.1). An arriving packet is discarded if there is no space in the queue. Uncompensated packet loss causes critical damage to voice quality for IP telephony. Each IP packet contains up to 60 ms of speech information. This upper bound matches the duration of critical units of speech called phonemes. Therefore, when a packet is lost, a phoneme may be lost in the continuous speech. The human brain is able to reconstruct some lost phonemes, but too much packet loss makes a voice unintelligible.

Besides, packet loss does not necessarily mean packets never arrived at their destination. For time-sensitive applications like voice telephony, a packet has to arrive in a certain time window. Otherwise they are dropped at the receiver and add to the packet loss. For non-real-time applications such as file transfers, packet loss is undesirable but not critical, since the protocols allow retransmission to recover dropped packets.

Packet loss can be avoided by traffic management mechanisms deployed in the IP network (see queuing delay in Section 2.2.1). QoS assurance in an enterprise network or single Internet service provider (ISP) environment is accomplished with little effort. A deeper problem is to administrate networks across multiple and independent domains with the current state of IP telephony.

Definitions of network packet loss and effective packet loss are given within TIPHON [154]. The IETF introduces several new metrics to capture packet loss patterns [138] within the IP performance metrics framework [133]. The ITU-T, finally, defines the so-called IP packet loss ratio as "the ratio of total lost IP packet outcomes to total transmitted IP packets in a population of interest" [116].

2.1.6 Measurement

This section concentrates on the determination of the voice quality, which has been introduced as voice-related parameters of the end-to-end QoS. Basically there are three key factors (clarity, delay, and echo) that must fall within certain bounds so that the quality of the received voice signal is judged acceptable. There is no absolute physical definition of voice quality. Two basic measurement techniques are available, which are both based on the subjective assessment of human listeners (see Figure 2.3) [146]:

Subjective methods have been used for a long time to determine the performance of new services. The objective of the real-time user assessment is to find out the average

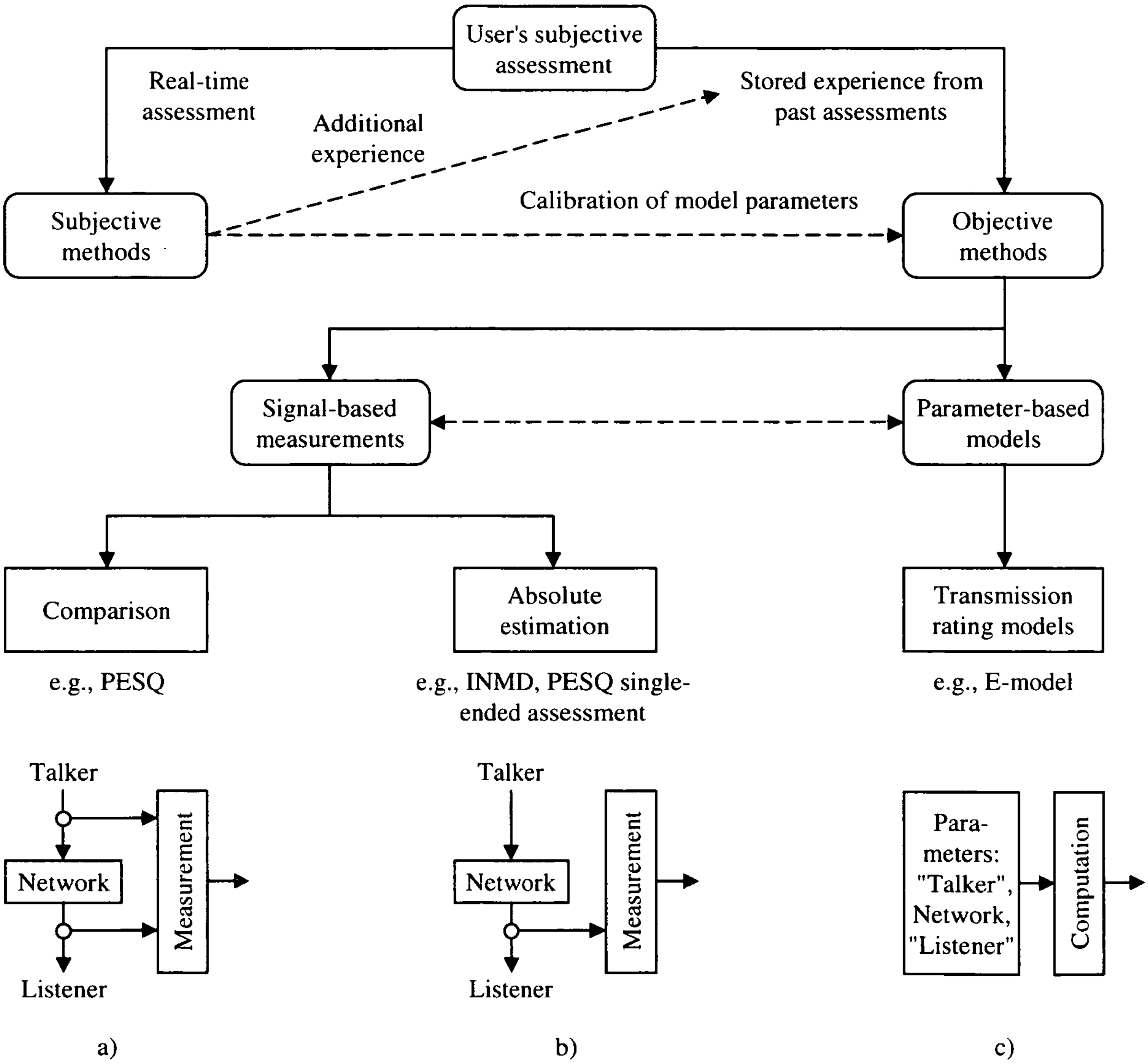
User's subjective assessment
Real-time assessment
Additional experience
Stored experience from past assessments
Calibration of model parameters
Subjective methods
Objective methods
Signal-based measurements
Parameter-based models
Comparison
Absolute estimation
Transmission rating models
e.g., PESQ
e.g., INMD, PESQ single-ended assessment
e.g., E-model
Talker
Network
Listener
Measurement
Talker
Network
Listener
Measurement
Para-meters: "Talker", Network, "Listener"
Computation
a)
b)
c)

Figure 2.3 – Classification of voice quality measurement methods: a) Comparison methods, b) Absolute estimation methods, c) Transmission rating models.

opinion of the end user.

- *Objective methods* have rationalized and supplemented some of the work necessary for subjective testing. Objective techniques use stored information on the user's assessment and, therefore, include some degree of calibration.

Subjective Methods

Subjective tests obtain the average user judgment of the voice quality. The subjects are asked a directed question and, in turn, provide a response out of a limited choice. A MOS is determined for a particular test condition by averaging the votes of all subjects. The relevance of MOS in terms of confidence intervals is determined by statistical significance analysis and reported along with the average values; the relevance is increased with the number of single votes. The results of subjective tests are influenced by a wide variety of conditions. Thus, the choice of the different parameters should be done very carefully. Generally, the most important of them are considered to be [146]:

- *Speech material*: Assessment depends on the gender of talkers, their pronunciation, the language, length and content of samples, the recording room and equipment characteristics.
- *Experimental setup*: Nationality and gender of listeners, recent previous experience with listening tests, instruction of listeners about the experiment, duration of test sessions, and order of presentation of speech samples have an influence on the rating.
- *Listening conditions*: Results depends on the loudness of the presented speech samples and the choice of equipment (headphones or telephone handsets).

ITU-T Recommendations P.800 [104] and P.830 [106] give guidelines on how to handle these factors to obtain reliable and replicable test results. Furthermore, subjective tests are categorized in these recommendations in four groups according to the number of talking participants. These measurement methods are also recommended for the assessment of network echo cancellers (see Section 2.3.7).

Conversational tests provide the closest simulation to real telephone conditions, because two subjects converse over a real connection. Such tests have pointed out that the system behavior under double talk influences the perceived voice quality to a great extent. That's why a specialization of conversational tests has been specified, the *double talk test*. Typically in such tests one subject is talking continuously while the other one is interrupting.

Talking and listening tests are designed for the evaluation of talking-related disturbances, such as impairments caused by echoes and switching. Only one subject is participating at the test; the other one is emulated according to the judged disturbance.

Listening-only tests play back pre-recorded speech material for evaluation purposes to the exclusively listening subjects. This method is well suited to compare the performance parameters of different terminals, algorithm implementations, or measurement conditions. Participants evaluate the quality either by using handsets in the case of a one-way communication or via correctly equalized headphones, when talking-related disturbances also belong to the subjects of interest. In the latter test subjects act as third-party listeners in the

Voice quality	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 2.1 – Listening quality scale for absolute category rating.

position of one of the two talkers and assess stereo samples recorded at artificial head measurement systems. The performance of the new approach on echo cancellation introduced within this work has been evaluated by such a third-party listening test. A detailed description is given in Section 2.3.7 and in Section 4. Listening-only tests are generally applied for the evaluation of specific parameters. The candidates are able to give a very detailed and precise description of the achieved transmission quality because they are "just" listening and do not have to concentrate on what to say.

Furthermore, listening-only tests differ in the presentation of the transferred (and mostly degraded) sample and, if included, the original voice stream. Namely, there are three techniques: the absolute, the degradation, and the comparison category rating.

The most frequently used subjective measurement method is *absolute category rating* (ACR). For ACR subjective tests, listeners are asked to rate the "absolute" quality of speech samples without comparison to a reference. The rating scale depends on the parameter to assess; a five-point scale recommended for the listening quality is given in Table 2.1 [104]. Generally, an ACR subjective test requires an average of 24 listeners [9].

ACR points out to be insensitive, when speech samples of good quality are evaluated (i.e., small differences in quality are not detected). In such cases *degradation category rating* (DCR) is better suited, where subjects listen to pairs (A–B) or repeated pairs (A–B–A–B) of speech samples. The sample A is a quality reference, and sample B represents a degraded version of A. The listeners are told to rate the samples B according to a degradation category scale listed in Table 2.2 [9]. Possible degradations to be evaluated are echo disturbances or speech gaps.

The results of DCR tests are reported in the form of degradation MOS (DMOS) and depend on the quality reference. Therefore, results of different experiments can only be compared if they rely on the same reference. In this work on a new approach on echo cancellation both methodologies—the ACR and a modified version of the DCR—have been utilized in a third-party listening test.

Finally, for the sake of completeness, *comparison category rating* (CCR) is mentioned. It is similar to DCR, but the order for reference sample and processed sample is random. Thus the second sample may be assessed better than the first one. This method is usually applied for the assessment of speech enhancement systems.

All in all, subjective tests are, especially in the case of VoIP, very informative, because they reproduce real service conditions that are experienced by final users. More exactly,

Voice quality	Score
Degradation is inaudible.	5
Degradation is audible but not annoying.	4
Degradation is slightly annoying.	3
Degradation is annoying.	2
Degradation is very annoying.	1

Table 2.2 – Opinion scale for degradation category rating.

auditive tests are able to evaluate echoes, noises, and delays; parameters that are hard to investigate with objective measurements methods. On the contrary subjective tests are time-consuming and costly. Furthermore, they are difficult to control, which means that consistency and repeatability are hard to obtain, and that they are not suited to investigate large parameter combinations. Because of these reasons much research and development effort has been put on objective measurement techniques in recent years.

Objective Methods

Objective measurement techniques are divided into three basic categories (see Figure 2.3):

- Comparison methods take into account both, the transmitted speech and the reference signal.
- Absolute estimation methods determine the absolute voice quality (i.e., the reference signal is not available for the algorithm). Both, the absolute estimation and the comparison method, belong to the group of signal-based measurements.
- Transmission rating models use a parameter-based model and compute the expected voice quality from data about the network; e.g. the E-model [65].

Basically there are no signal-based measurement methods available for the evaluation of the subjective perception of the conversational speech quality. Such a technique would have to provide a psychoacoustic model that includes all talking and listening-related attributes of a conversation—mainly the speech clarity, delay, and echo. The E-model also discussed in this clause provides only a rough a-priori estimate under special circumstances to evaluate the conversational quality and is not capable to assess occurring signals on the network. That's why the focus is here on the objective determination of the speech clarity, which has been introduced as the one-way measure of the conversational voice quality. For the measurement of delay and echo the reader is referred to the corresponding Sections 2.2 and 2.3, respectively.

There are different reasons why it is important to quantify voice quality in an objective manner [55]: The carrier or service provider wants to compare the quality of the VoIP network to the PSTN, because the latter has become the standard for acceptable voice quality. Additionally, system engineers need to determine the effects on voice quality, when they change the design or vary the network conditions. Lastly, objective measurement of voice quality allows service providers to compare their offerings with each other and, moreover, the results may serve as a basis for the voice quality service level agreements.

Modern digital compression techniques require more sophisticated objective measurement methods than just the determination, for example, of signal-to-noise ratio (SNR) and total harmonic distortion (THD) of the transferred signal. The implemented coding algorithms make use of the human physiology and perception to hide quantization distortion in areas of the signal that humans cannot perceive. The engineering tools also have to consider the human auditory system to provide reliable results over a broad range of applications and should, moreover, provide network technology independent assessment.

Objective testing techniques evaluate attributes of the received voice signal with the intention to estimate the user's assessment of the perceived voice quality. The different parameters of the applied algorithms are calibrated, consequently, in the design stage of the objective method according to a large database containing results of subjective tests.

Generally there are two means on how to provide the test signal necessary for the evaluation of the voice quality: intrusive and non-intrusive. When explicit test traffic is injected into the system, an *intrusive* (also called direct, explicit or active) measurement is applied. It comes with the advantage that the signal characteristics are given by the measurement system and are, hence, chosen by the system engineers (also see delay measurement methods in Section 2.2.3). Furthermore, different implementations are compared with each other in a more reliable way when utilizing standardized speech samples. Besides, real voice samples provide more realistic test conditions. The drawback of an intrusive measurement is that the additional traffic may influence the system behavior and that such an assessment may only be feasible when the system is offline. To overcome these shortcomings, measurement methods evaluating the life traffic have been developed; they are called *non-intrusive*, implicit or passive assessment techniques. Such methods make a larger number of tests possible, and they are less costly, because life traffic sampling is feasible. On the contrary they offer lower accuracies than intrusive techniques and are more difficult to develop. Especially in the field of perceived voice quality measurement much research work has been done and is still going on.

Generally, objective methods rely on one of two distinct and complementary techniques (see Figure 2.3):

- Parameter-based methods involve models of the transmission system.
- Signal-based methods are either relying on the voice sample after transmission or they use both the original signal and the transferred and—with the exception of voice enhancement systems—degraded signal.

Parameter-based methods are chiefly represented by the E-model, which was originally developed in ETSI for the needs of network planning. The E-model computes and outputs the transmission rating factor in ISDN, PSTN and IP telephony systems as estimation for the expected conversational voice quality. The calculation is exclusively based on parameters of the network and the talker's and listener's conditions and there is no a-priori knowledge about the characteristics of the applied signals. Examples for possible parameters are the end-to-end delay, different codec implementations, and packet loss in an IP environment. The algorithm of the E-model is determined in [65]; an application guide is given by [66].

Signal-based techniques, which rely on the transmitted voice stream and do not take into account the original voice stream, are called absolute estimation methods; thus they are applicable for non-intrusive measurements. The in-service, non-intrusive measurement device (INMD [102], for example, measures the conversational voice quality in terms of parameters such as speech level, noise level, echo loss and echo path delay. It does not provide an estimate of the perception of this conversational quality, as psychoacoustic effects are not considered. To overcome this lack, the ITU-T Study Group 12 is currently conducting a competition with the intention to bring out a standardized method as an extension to perceptual evaluation of speech quality (PESQ) for the so-called single-ended assessment. PESQ is a comparison based method for the evaluation of the speech clarity.

Signal-based comparison methods input both, the reference and the transmitted signal received at the listener's end of the one-way communication, to the clarity algorithm. Typically tests are carried out using speech from both male and female speakers at a standardized active speech level around -18 dB_{m0}. The utterances of the samples are carefully chosen to provide the system with a comprehensive range of speech sound patterns [55]. Furthermore, the algorithms are calibrated for a limited range of languages; in most cases for German, English and French, but they may also be applied for similar languages like Italian or Spanish.

All algorithm comparison techniques contain some basic functions applied on both signals before a MOS-related value is presented at the output [55] [59]: As a first step, the reference and the degraded signal are time-aligned. The implemented de-jitter buffers in packet networks compensate varying delays, but may also use heuristics to dynamically adjust their buffer lengths in silent periods. However, the voice stream may experience varying end-to-end delays, which makes the time-alignment much more difficult. Second, the reference and received signals are brought on the same power level by gain-scaling. The problem arises, when some impairment temporarily alters the received signal level. Therefore, the same level is only achieved in an average sense. Afterwards both signals are transformed to the frequency domain and the resultant spectrum is assigned to bands (also called bins) that represent the non-linearity of the human ear. As a next step the contents of each bin of the reference and the transmitted signal are compared; a cognitive model is applied to evaluate the significance of the differences (disturbance processing). The result of this comparison is a clarity score for each part of the utterance. Finally an aggregated score for the whole voice sample is calculated and, in most cases, an estimation of the MOS value is built. The quality of the algorithms is determined by correlating the objective with the subjective MOS results of the same samples. High correlation larger than 0.9 should be achieved.

The three most significant algorithms for the measurement of the perceived voice quality are:

- The perceptual speech quality measure (PSQM) is the original ITU-T standard for intrusive assessment of the perceived voice quality [109]. It was developed by KPN and was adopted by ITU-T in 1996. However, the scope of the standard is limited to the assessment of codecs; networks cannot be evaluated. PSQM has been replaced by PESQ in February 2001.

- The perceptual analysis measurement system (PAMS) was developed by British Telecom [60]. It was published in 1998 and it was the first method to provide robust assessment in the case of varying delays (e.g., VoIP).
- The PESQ is the current ITU-T Recommendation for signal-based comparison [110]. The ITU-T carried out a competition in 2000 with the intention to find out an objective method for a broader range of applications. The participants of the contest had to apply their objective algorithm on various databases obtained in subjective listening tests and the correlation factor was taken as a measure. The output of this contest was a new standard built on both the PSQM and PAMS. PESQ is able to evaluate a larger number of codecs than the two previously mentioned methods; even transcoding is assessable. Furthermore, it is able to cope with varying delays and packet loss, which is especially important for VoIP systems. Moreover voice streams with silence substitution and transcoding deliver reliable results. PESQ is not applicable for the assessment of one-way delay, non-intrusive and broadband systems, sidetone and talker echo, acoustic interfaces as well as for vocoder with bit rates below 4 kbit/s. Current standardization efforts focus on acoustic interfaces and single ended (i.e., non-intrusive) assessment.

Objective methods are repeatable, efficient and, in comparison to subjective testing, fast. Therefore, they are ideally suited for testing large numbers of parameter combinations.

2.1.7 Classification

ETSI TIPHON was one of the earliest standards bodies to define different levels of end-to-end quality categories for IP telephony networks. As the three classes incorporate the end to-end delay, they represent a restricted version of the conversational voice quality (e.g., the talker echo is not considered). The network and the terminal characteristics are considered in this classification [153]:

The speech clarity describes a relative measure of the listening quality in terms of a certain codec. The absolute quality of this measure depends on the individual terminal as

	Wideband	Narrowband			Best Effort (Note 2)
		High	Medium	Acceptable	
Speech clarity	better than G.711	equivalent or better than G.726 at 32 kbit/s [86]	equivalent or better than GSM-FR [149]	Not defined	Not defined
End-to-end delay	< 100 ms	< 100 ms	< 150 ms	< 400 ms	< 400 ms
Transmission rating factor R	Note 1	> 80	> 70	> 50	> 50

Table 2.3 – ETSI TIPHON end-to-end quality categories for VoIP services. Note 1: The R-value characterization within the E-model for wideband codecs is under study. Note 2: The delay and the R-value for best effort class are target values and, therefore, not guaranteed.

well as on the quality of the network to be assessed. The transmission rating factor is computed according to the E-model (see Section 2.1.6). It depends on various parameters of the VoIP system like the implemented codec and the end-to-end delay (an example of the impact of the end-to-end delay on the overall quality computed by means of the E-model is given in Section 2.2.2). Furthermore, a typical situation using a "standard" telephony handset according to ITU-T Recommendation P.310 [98] is assumed. The three performance parameters of Table 2.3 are measured according to the corresponding TIPHON document [154].

The TIPHON mouth-to-ear quality classes wideband and narrowband are implemented over QoS-engineered IP networks and both provide performance guarantees for 95 percent of all connections; the best effort class gives no QoS guarantees. The narrowband class is subdivided into three categories: The high class represents the quality level of recent ISDN services. The medium class is equivalent to recent wireless mobile telephony services in good radio conditions. Finally, the acceptable class corresponds to common wireless mobile telephony services.

2.2 Delay

> Delay is defined as "the time between the instant at which a given event occurs and the instant at which a related aspect of that event occurs."[16]

2.2.1 Sources

Delay is one of the biggest technical challenges in IP telephony. It is also one of the main contributing parameters for the annoyance of echo (see Section 2.3.3), which is seen as another key factor influencing the conversational voice quality. The one-way delay (also called end-to-end, user-to-user, transit, or transfer delay; or in an acoustic environment defined as mouth-to-ear delay) results from the sum of various delay-introducing components along a signal path. It is chiefly experienced in the IP network or at the interfaces to other communication networks. A general example with different IP endpoints of a communication, such as a multimedia personal computer (PC), an IP phone, and an IP gateway, is illustrated in Figure 2.4 [56]. Other configurations are easily deduced from this arrangement.

The corresponding portions in Figure 2.4 are either fixed and known in advance or they depend on the current state of the system and are thus variable and unpredictable. Each of the delay components in Figure 2.4 is explained below, starting at the left hand side of the illustration with the time taken by the multimedia PC and proceeding with the succeeding components in Figure 2.4.

---

[16] Glossary of Telecommunication Terms-Federal Standard 1037C.

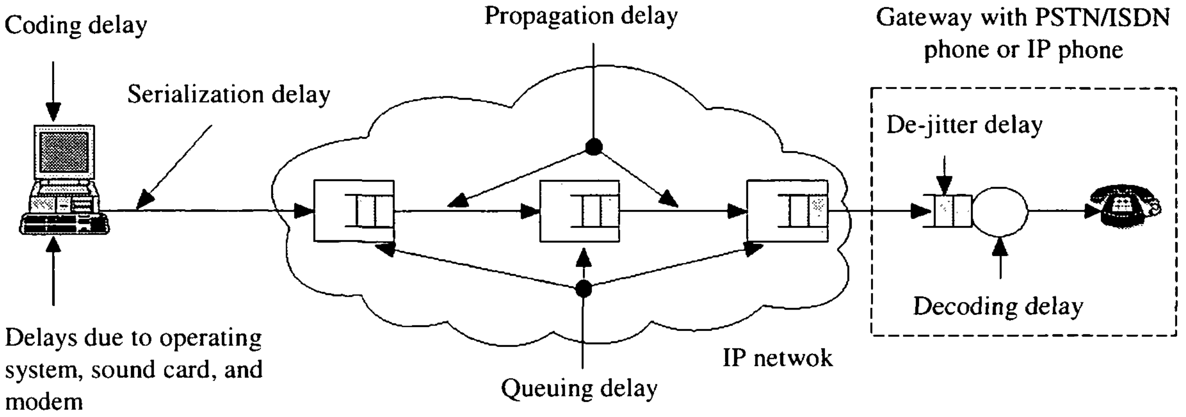

Figure 2.4 – Delay components in an IP telephony environment.

Operating System Delay

Beside delays caused by modems in dialup networks, packet voice systems using multimedia PCs are responsible for delays due to operating system inefficiencies and sound card delays. The last two time spans can be minimized by using a gateway card that is based on a fast, specialized digital signal processor (DSP). A PC sound-card, for example, typically introduces 20 ms of output delay.

Most IP phone applications are regular programs running on top of an operating system such as Windows. They access sound peripherals through an application programming interface (API, e.g. Wave API for Windows) and they send voice packets to and retrieve them from the network via the socket API. The sound card samples the signals from the microphone and accumulates the samples in a buffer. Having filled this memory buffer, it triggers an Interrupt to give notice to the operating system. If there are enough samples in the buffer to build a full frame for the compression algorithm, the samples are put on the network by means of the socket API. As most operating systems can only deal with a certain amount of Interrupts per second, a minimal accumulation delay is introduced by the operating system. On a Windows system the samples are available in time gaps of more than 60 ms [8], independent of the codec used by the program. The same situation occurs when playing back the samples, and there are further delays introduced by the socket implementation. As the time span from triggering an Interrupt request until the operating system carries out the Interrupt is unforeseeable (it is only limited by an upper time boundary), the operating system delay is variable and unpredictable.

To sidestep this limitation, IP telephony gateways and IP phones are based on a real-time operating system. Another way to cope with this drawback is to carry out all the real-time functions (sample acquisition, compression and RTP) by selected hardware, and execute only the control functions from the non real-time operating system.

Coding and Decoding Delay

As already mentioned in Section 2.1.3, most codecs produce frames of a fixed length (the frame length or frame size). One or several frames are grouped together and sent in one IP packet. The audio stream needs to be accumulated first before the coding algorithm is ap-

Codec	1st introduced in the year	Algorithm	Bit rate	Frame size	Look-ahead	Silence compression		MOS
			[kbit/s]	[ms]	[ms]	DTX/VAD	CNG	
G.711 [82]	1965	PCM	64	0.125	0	–	App. II [83]	4.2
G.723.1 [84]	1996	MP-MLQ	6.4	30	7.5	G.723.1 [84]	Ann. A [85]	3.9
		ACELP	5.3					3.7
G.729 [87]	1996	CS-ACELP	8	10	5	Annex B [88]		4.0

Table 2.4 – Key attributes of the G.711, G.723.1, and G.729 codec.

plied and the coded speech frame is generated. That's why there is a delay of one frame before processing takes place. In addition to that many coders also use some data of the succeeding frame to improve compression efficiency. The time taken by the coder for this measure is named the *look-ahead delay*. Three commonly used codecs and their main attributes are listed in Table 2.4.

When assuming an efficient use of the processor resources, the time necessary to compress one input frame should be the same as the frame length. Thus, the *frame processing delay*, which is referred to as the time required for accumulating and processing the speech frame, is not larger than twice the frame size and it does not include the look-ahead time. With higher compression rates more computation power has to be reserved for the selected algorithm. The sum of the frame processing and the look-ahead delay is defined as *coding delay* (see Figure 2.4). Some references exclude the time needed for accumulating the data (which equals the frame size) from the frame processing delay [56]. In this case the term *algorithmic delay* is sometimes introduced for the sum of the delay for processing one single frame and the look-ahead delay (instead of coding delay). The decoding at the receiver takes less time than the coding at the sender. Some guidelines denote that the decoding delay is normally half the algorithmic delay [56]. Other definitions of the coding delay also include the time to put the coded data on the transmission medium [70]. This serialization delay (see below) is clearly separated from the coding delay in this work. The example in Figure 2.5 shows the composition of the overall coding delay and the maximum serialization delay in an IP network with two voice frames in one IP packet [70].

The coding delay in IP-based systems with N frames per packet equals

$$(\mathrm{N}+1) \times \text{Frame size} + \text{Look-ahead}\,, \tag{2.1}$$

and is easily computed for the three frequently implemented coders with the values listed in Table 2.4. Moreover an extensive variety of coders for IP-based applications is given in [70]. In order to reduce one-way delays, the codec should have a short frame length. On the contrary, coders with larger frames tend to be more efficient and have better compression rates. Another point is that a 40 byte overhead is added by the transport protocols (RTP, UDP and IP packetization, see Section 2.1.2) for each packet transmitted through the network. If each compressed voice frame is transmitted in a packet of its own, for some

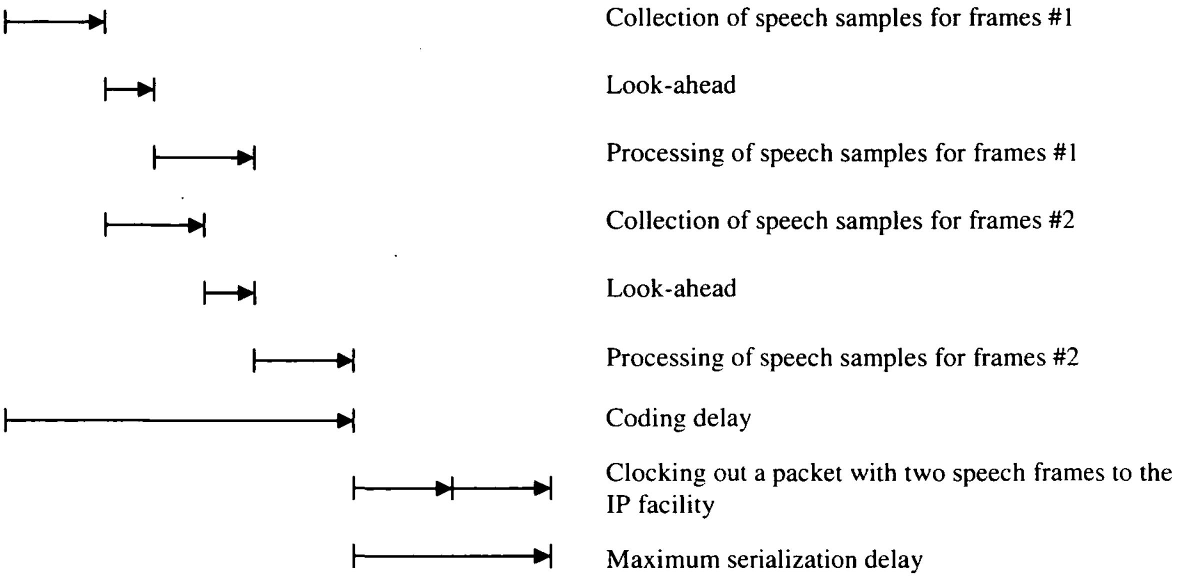
Collection of speech samples for frames #1
Look-ahead
Processing of speech samples for frames #1
Collection of speech samples for frames #2
Look-ahead
Processing of speech samples for frames #2
Coding delay
Clocking out a packet with two speech frames to the IP facility
Maximum serialization delay

Figure 2.5 – Portions of the total coding delay and the serialization delay for two audio frames in one IP voice packet.

coders, the overhead will be comparable to if not greater than the useful data. In this case the bandwidth is not utilized very efficiently, which is also often referred to as poor *packetization efficiency.*

Alternatively, the required bandwidth for one frame per packet can be reduced when deploying the *RTP header compression*, which reduces the overhead from 40 bytes to 2 to 4 bytes [136]. A 2 byte header can be achieved when the UDP checksum is not used and a 4 byte header when it is used. The RTP compression scheme utilizes the nature of UDP, as most of the bytes in the UDP headers of consecutive packets remain unchanged during a connection [10].

A common way to decrease the consumed bit rate is to place two or more frames in one voice packet (see Figure 2.6). Sometimes the term *packetization delay* is introduced for this time span, which is understood as the time needed for the packing of the extra frames, or as the whole delay from the sampling to the start of the sending process (i.e., the coding delay), or just as the sum of the frame lengths required for accumulating the voice samples of one packet [39].

If all frames accumulated in the packet belong to the same audio stream, the coding delay will be increased according to Equ. (2.1). In this case it would be more efficient to use a coder with a longer frame size. Typically packets contain between 10 to 30 ms of speech (out of one or more frames) which provides a practical trade-off between network efficiency and increased delay [58]. An alternative way of packing multiple frames per packet without any impact on delay is to put frames of different audio streams, but with the same network destination, in one packet. This is often the case for VoIP trunks between corporate sites, or between gateways inside a VoIP network.

Another delay parameter that has to be taken into account before the packets are sent is the *redundancy* policy. Techniques such as forward error correction (FEC) are available to

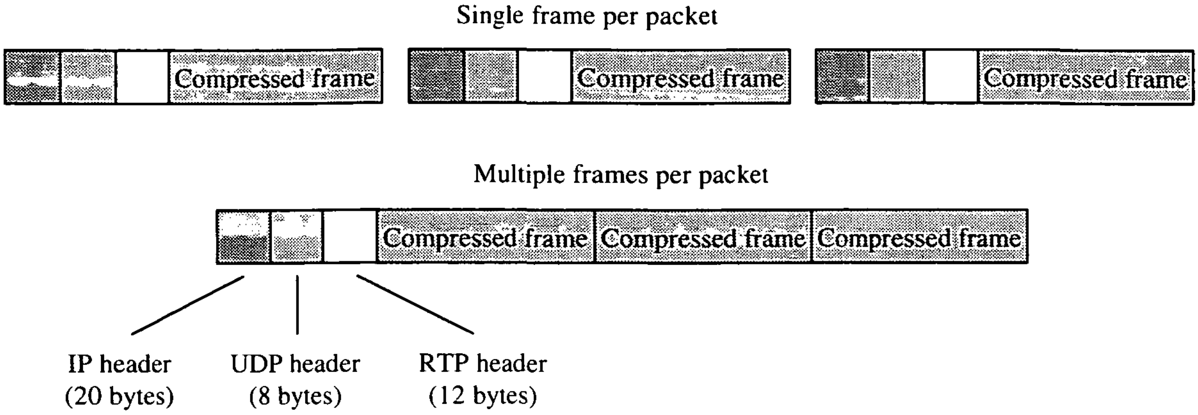

Figure 2.6 – Improving packetization efficiency with multiple frames per packet instead of one frame per packet.

lower the frame loss rate in terms of a given packet loss rate. To accomplish this, redundant information is spread over several packets so that the frame information can be recovered even if some packets are lost. The drawback of a FEC system is its impact on the consumed bandwidth and especially on the delay. Therefore, packet or frame loss is mostly handled by the receiver and not by the sender with PLC systems (see Section 2.1.3).

Serialization Delay

The serialization delay (sometimes called transmission delay) is the time it takes to put the whole IP packet on the transmission media. It is determined by the size of the IP packet and the available bit rate on the line:

$$\text{Serialization delay} = \frac{\text{Packet size}[\text{bits}]}{\text{Bit rate}[\text{bits/s}]}. \tag{2.2}$$

In other words, the higher the line speed and the smaller the packets, the smaller is the serialization delay. The accessible amount of transmission resources for the IP packets depends on the traffic load on the access link to and on the medium itself. On the average, the compressed frames should at least be clocked out with the same rate the samples are collected at the input of the encoder. Otherwise, when the sending buffer is filled with frames to be transmitted, speech frames are dropped at the sender. The serialization delay should not, in case of multiple frames per packet, exceed the length of the frames contained in one packet:

$$0 \leq \text{Serialization delay}[\text{s}] \leq \text{Frame size}[\text{s}]. \tag{2.3}$$

The maximum transmission delay is also depicted in Figure 2.5.

Queuing Delay

Queuing delays are observed, when IP packets are going through transmission and switching points of the network (e.g., gateways and routers). The voice packets arriving from different links are queued into buffers for each output link where they wait until they are sent out on that link. The number of packets in the buffer and the transmission capacity of the corresponding outgoing link determine the queuing delay. When the buffer is filled up, subsequent packets arriving for that certain link are discarded. As the current traffic load is statistically distributed, the queuing delay is variable and unpredictable. One measure to decrease this delay is to deploy faster transmission links (overprovisioning), but this is only feasible, when the network designer has control of the infrastructure such as in corporate IP networks. Besides, the Internet Engineering Task Force (IETF) defines explicit resource management schemes for time-sensitive traffic. The most popular among these are the Diffserv [135] and the Intserv [128] approach; both describe network architectures. Furthermore, the IETF standardized the resource reservation protocol (RSVP) [132] for signaling purposes, which was originally dedicated to the Intserv model and extended later for other deployments. Finally the multi protocol label switched paths (MPLS) architecture [142] has been added to the so-called IP QoS architecture within the IETF. The different schemes reserve bandwidth capacities and prioritize marked IP packets to minimize queuing delays for voice and other delay-sensitive applications.

Propagation Delay

This portion of the end-to-end delay (also called transmission delay, see serialization delay) and denotes the time the voice packets need to pass the whole path via the transmission links (without the queuing delay, which is encountered in the nodes of the network). It is determined by the speed of light on the medium and by the distance the packets are traveling on that link:

$$\text{Propagation delay}\,[\text{s}] = \frac{\text{Distance}\,[\text{m}]}{\text{Speed of light}\,[\text{m/s}]}. \tag{2.4}$$

Planning values for the delay of various transmission elements are given in [70]. In most cases the resulting delay values of Equ. (2.4) are fixed and they become relevant, when long distances are involved. This happens, for example, when using a geostationary earth orbit (GEO) satellite, which is one with a constant position with respect to earth and with an orbit about 36000 kilometers above the earth's surface [55]. This distance corresponds to a round-trip delay of around 500 ms, when assuming in Equ. (2.4) for the speed of light values close to vacuum conditions (3.10^8 m/s). These delay values may be reduced by using low earth orbit (LEO) satellites, where connections from earth may be handed over from satellite to satellite. This can occur, because LEO satellites are moving with respect to the ground station. On the one hand, the delays are reduced compared to GEO systems. On the other hand, LEO satellites introduce variable delay paths and unforeseeable buffering times during connection handover. In the case of a public land mobile network (PLMN) the propagation and also the overall one-way delay are quite constant and predictable. The variation of the involved distances and of the corresponding delays originated by handovers from one cell to another can be neglected. In IP telephony networks the voice packets of one media stream may traverse different transmission links, which in most cases has the consequence of differing distances. Therefore, long distance calls may also come with varying propagation times. Another term, often found in literature, is the *network delay*, which comprises the serialization, the queuing, and the propagation delay. That's why it is understood as the time from the first bit put on the network until the last bit arrives at the receiver [39].

De-Jitter Buffer Delay

Generally, variable delays in packetized transmission systems are called *jitter* (or packet delay variation). Jitter occurs due to the varying queuing delays in the network and changing propagation delays in the case of connections over LEO satellites. Furthermore, IP packets of the same stream may even take different paths in the IP network and experience different delays in this way. And, in the end, every operating system comes with some amount of variable and unpredictable delay (see operating system delay above). But this last effect can be neglected, when using selected hardware with real-time operating systems. Hence, this unavoidable effect depends strongly on the specific mechanisms for transport, queuing and prioritization (see queuing delay in this section), which may be implemented in such a system. Network jitter may dominate even for low average network delays. If an IP packet does not arrive in time at the receiver, it is discarded and adds to the packet loss. If this happens too often, the quality of voice will be affected significantly. Definitions of the term jitter are either statistically or they are given for each pair of packets:

- ETSI distinguishes between the two-point and the one-point packet delay variation [154], which, in both cases, denotes the difference between upper and lower percentiles of the packet delay distribution. The two-point packet delay variation uses two monitoring points, whereas the one-point relies on only one point at the receiver. The two-point measure exploits the difference between the inter-packet sending and inter-packet arrival times and it is measured with two synchronized test boxes. The second one is only based on the inter-packet arrival times. Moreover it requires only one test box at the receiver, and, therefore, no synchronization process is needed. The one-point packet delay variation represents the end-systems view of jitter, but is not capable to localize and quantify the sources of the packet delay variation in the network.
- The IETF defines the IP packet delay variation as follows [145]: "A definition of the IP packet delay variation (IPDV) can be given for packets inside a stream of packets: The IPDV of a pair of packets within a stream of packets is defined for a selected pair of packets in the stream going from measurement point 1 to measurement point 2. The IPDV is the difference between the one-way-delay of the selected packets." This is equivalent to the RTP definition of jitter [129].

Delay variation must be compensated before playing the received voice samples to the user. Otherwise significant voice quality degradations would be perceived. The compensation is realized on the application layer of the system at the receiving side by collecting the packets in a *de-jitter buffer* (also called jitter buffer or play-out buffer). This buffer rearranges the timely order of the packets by holding the packets for a while in order to synchronize voice samples of different IP packets before playing them out. The *de-jitter delay* (also called holding time or play-out delay, see Figure 2.4) corresponds to the time packets stay in the buffer, which is less than the buffer size. The time of departure of each packet is

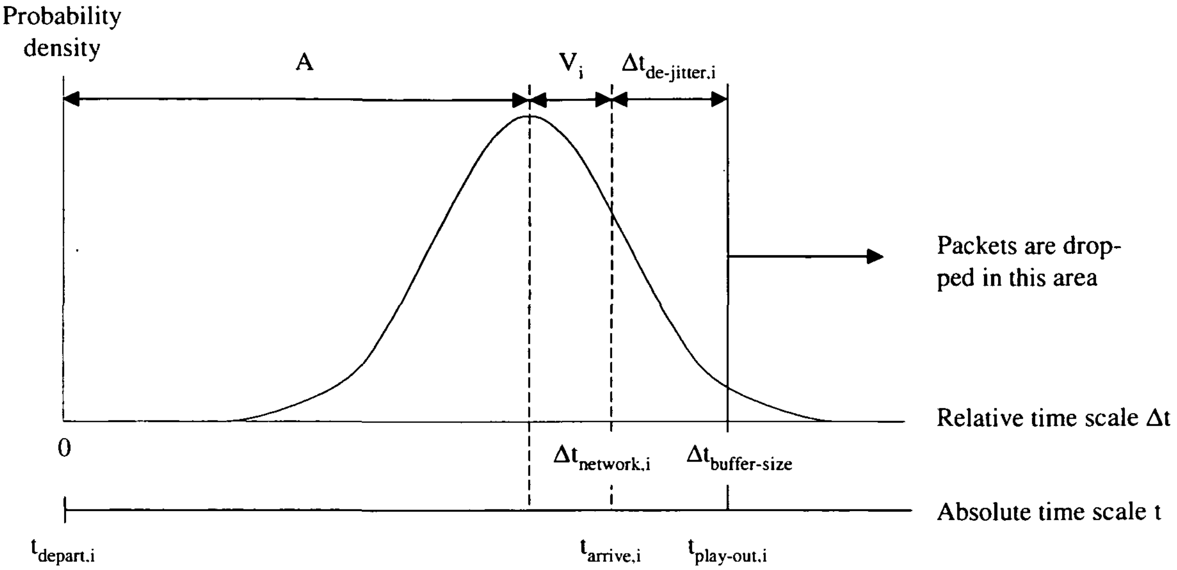

Figure 2.7 – An example for the distribution of delay values of an IP stream. The average delay of all values (A), the variable (V_i) part, and the de-jitter buffer delay of the i-th packet is pointed out in order to explain the de-jitter buffer performance.

determined by reading out the timestamp information provided by RTP (also see Section 2.1.2).

An example illustrating the de-jitter buffer behavior is given in Figure 2.7. The i-th voice packet, departing from the sender at the time instant $t_{depart,i}$ and arriving at the ingress of the de-jitter buffer at $t_{arrive,i}$, experiences the following delay over the IP network:

$$\Delta t_{network,i} = t_{arrive,i} - t_{depart,i}, \tag{2.5}$$

which equals the network delay minus the serialization delay according to the definitions introduced above. Figure 2.7 contains an example of network delay values, an approximated distribution function of those values, and the distinct portions of one IP packet as an example. The delay value for this packet can be divided into an average part A calculated from all single values—characterizing the sum of the propagation delays and the mean value of the queuing delays—and a variable part V_i. The de-jitter delay (i.e., time of the packet in the de-jitter buffer) of the i-th packet depends on the play-out time $t_{play\text{-}out,i}$:

$$\Delta t_{de\text{-}jitter,i} = t_{play\text{-}out,i} - t_{arrive,i}. \tag{2.6}$$

Packets arriving after the play-out time at the sender ($t_{arrive,i} > t_{play\text{-}out,i}$) are discarded; the amount of speech in this certain packet is lost. This adds to the packet loss and impairs the voice quality. Other portions of the overall one-way delay, like operating system delay, may also induce varying delays. As jitter is mainly assigned to networks they are not considered here.

The length or size of the de-jitter buffer $\Delta t_{buffer\text{-}size}$ corresponds to the longest tolerable delay of the packets through the network. Packets that experience the minimum delay until they reach the receiver will spend the maximum time in the de-jitter buffer before being played out as a synchronous stream and vice versa (i.e., packets with the maximum delay stay the minimum time in the queue). According to the ITU-T, the de-jitter buffer should add only one half of the buffer size to the mean network delay [70]. A de-jitter buffer, for example, with a range for a maximum packet delay variation of 50 ms, should introduce 25 ms additional delay on average. If there is high jitter on the network, the overall one-way delay is high because of the de-jitter delay even if the average delay is low. The selection of the jitter buffer size is very critical to IP based telephony systems. An optimum buffer size has to be established which balances the remove of jitter and the limitation of delay to tolerable levels. If the buffer is set too low, some packets may be lost. If it is adjusted too high, higher delays are the consequence. Basically play-out buffers are categorized into two types [58]:

Static de-jitter buffers have a fixed size buffer length and are, thus, easy to implement and manage. The buffer size is manually configurable. Generally a static buffer should be based on a well managed underlying network to keep jitter within the size of the buffer.

The more advanced dynamic de-jitter buffer adjusts its play-out point according to the history of the arriving packet's jitter values. This suits to systems with more varying network conditions and, above all, unpredictable jitter profile. A well designed adaptation algorithm helps to decrease the end-to-end delay and thus improves the perceived conversational quality experienced by the customers. The jitter buffer size may be determined by using the ratio of late packets to those that arrive in time. Four basic playout mechanisms for audio applications in WANs are given in [57]. Initially, these heuristics may take some adaptation time, because the control unit needs to determine the current jitter in the network. A basic approach for such algorithms is, for example, to start with a very small buffer size, and continually increase it until the average percentage of packets arriving too late is below one percent [8]. A dynamic play-out buffer may also be useful on a well-managed network. If such a network offers a better performance than specified, e.g. below 10 ms of inter-packet arrival time rather than below the planned and expected value of 30 ms, the buffer is adapting to decrease the overall delay of the connection. The time, when to adjust the play-out point of the dynamic de-jitter buffer, represents an important design consideration. It is, in some cases, possible to adapt during active speech periods without being noticed by the user. Nevertheless, it is much better, to accomplish this adjustment during silence, where it is almost impossible to perceive.

2.2.2 Interactivity

When the end-to-end delay experienced by the user exceeds a certain limit, the conversation tends to a half-duplex mode (i.e., conversation is only possible in one direction at a time). This restricted mode of conversation occurs due to the following problem for the users: The larger the end-to-end delay, the longer have the participants to wait before the other one reacts. If there is not enough patience for the large delay, the users go on talking before the reply of the other party has arrived. When this happens, the talkers interrupt and, thus, disturb each other.

The impact of the mouth-to-ear delay on the perceived voice quality of a standard connection can be shown with the E-model (see Figure 2.8) [70]. The graph of the transmis-

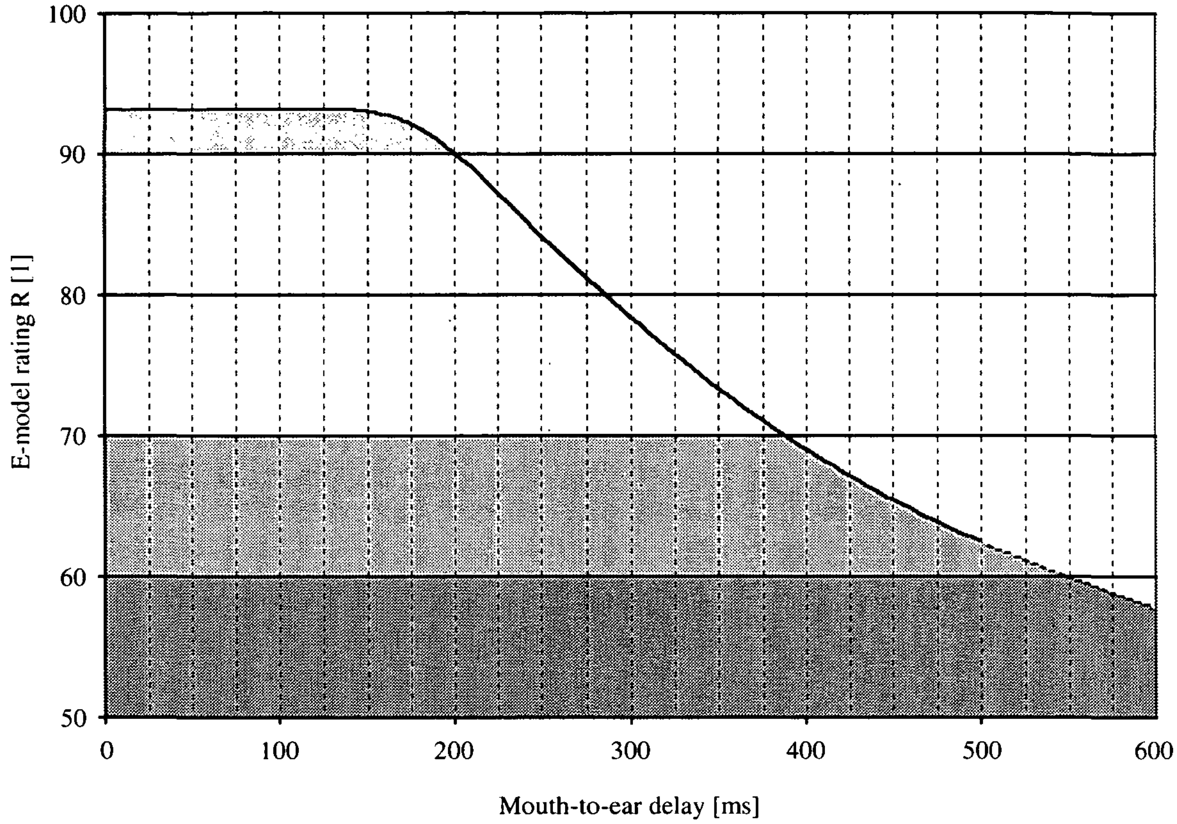

Figure 2.8 – Determination of the effects of the end-to-end delay on the conversational quality calculated by the E-model.

sion rating factor R starts to decline at about 150 ms; with highly interactive tasks (e.g., some speech, video conferencing and interactive data applications) it may even be affected by one-way delays below 100 ms. Echoes and degradations caused by voice coding or processing are not taken into account, i.e. the diagram represents a pure and undisturbed connection with varying end-to-end delays. Delay values above 500 ms are represented by a dashed line to indicate that these results are not fully validated. But it represents the best estimate of the expected rating factor, and, therefore, provides a useful guideline.

2.2.3 Delay Measurement

The delay performance of an IP telephony system may vary considerably during a call. Therefore, it is essential to emulate a broad range of different network conditions in order to derive relevant statistics. The collected data are generally evaluated in terms of the average delay, the standard deviation, and the overall range of the delay. Possible uncertainties stem from the varying network conditions (e.g., varying loads on the IP links), the scheduling mechanism of the operating system, and the convergence procedures due to the many heuristics implemented in an IP based telephony system (e.g., dynamic de-jitter buffer). These points have already been discussed comprehensively in Section 2.2.1. Because of the heuristics it is necessary to allow a short convergence time before the assessment starts.

As already mentioned in Section 2.1.6, there are two means on how to provide the test signal for delay measurement: it is either injected by the measurement device or the live traffic is evaluated, which refers to as intrusive or non-intrusive measurement. In the case of an analog interface for mouth-to-ear delay measurement intrusive techniques are preferred. The attributes of the artificial and, thus, non-speech like test signal are chosen by the system engineer. When deploying, for example, a reference signal with a distinct peak in its auto-correlation function, the original and the received signal are easily time-aligned by means of cross-correlation in order to determine the wanted delay. This method has been chosen for the Agilent VQT (Voice Quality Tester) [168]. Non-intrusive techniques are applied for pure IP measurements, because the timestamps of the transferred packets are independent of the payload.

The statistical evaluation uses several results of single measurements on the transmission layer. Principally there are two delay measures, the round-trip or the more complex one-way delay.

Round-Trip Delay

The IETF has established the IP performance metrics (IPPM) working group [165] to approve guidelines for measuring delay, delay variation, and loss in an IP environment. The determination of each metric is based on the framework for IPPM [133], which includes some basic definitions. One of these metrics is the *round-trip delay*, which is defined as follows [139]:

> "A type-P-round-trip-delay from Src to Dst at T is ΔT means that Src sent the first bit of a type-P packet to Dst at wire-time T, that Dst received that packet, then immediately sent a type-P packet back to Src, and that Src received the last bit of that packet at wire-time T + ΔT."

The Src (from "source") denotes the IP address of the host at the beginning of the IP

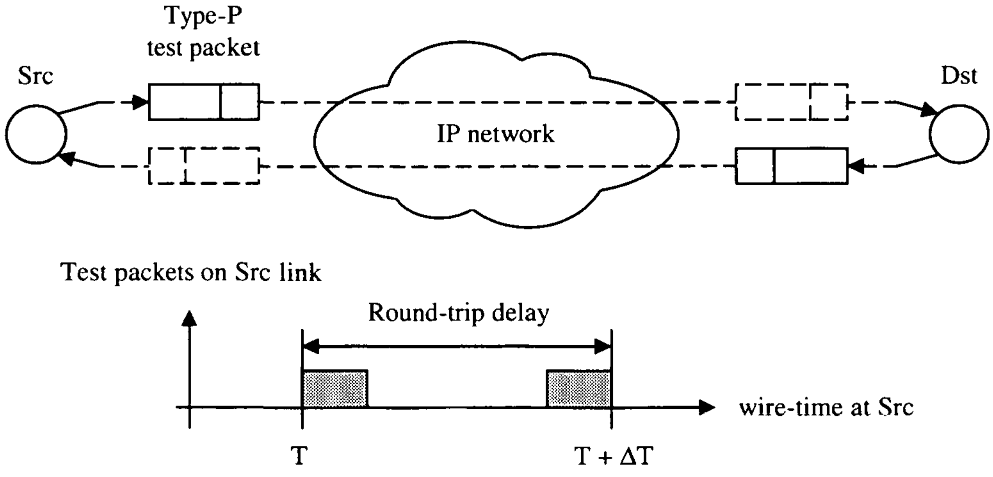

Figure 2.9 – Measurement of the round-trip delay according to IETF: One test packet is transmitted from Src to Dst, and another one from Dst to Src immediately after the receipt of the first. The round-trip delay corresponds to the difference in the wire-times.

path through the Internet, and Dst (from "destination") represents the IP address of the host at the end of that path (see Figure 2.9).

The wire-time represents the time at which the bits appear on the observational link of the hosts. The wire-times in the definition above identify two timely instances [139]: First, when the test packet leaves the network interface of Src (at wire-time T) and, second, when the corresponding reply packet completely arrives at the network interface of Src (at wire-time T + ΔT). If the timings are provided by software on Src, then this software measures the time gap from the moment, when Src grabs a timestamp just prior to sending the test packet, and when it takes a timestamp just after having received the response packet. The value of type-P-round-trip-delay depends on the type of protocol ("type-P", e.g., UDP [125] or TCP [124]), the port number, the size, or the arrangement for special treatment (e.g., IP precedence or RSVP [129]. The exact type-P used to make the measurements must be accurately reported. The ping program, for example, is a widely available application for round-trip delay determination. The type-P is of Internet control message protocol (ICMP) echo request and reply with a packet length of 60 bytes.

The methodology of measurement, for a given type-P, starts with the formation of a *test packet.* It includes the Src and Dst IP address as well as a padding portion, which should be filled with randomized bits to avoid unpredictable amount of compression along the path, which would alter the perceived delay. The test packet must contain some identifying information so that the response to it can be identified by Src when Src receives the response.

The description of a certain measurement method for the determination of the round-trip delay should include an analysis of various *errors* or *uncertainties* occurring due to:

- The difference between wire-time and host-time.
- The uncertainty in the *clock* of the Src host. Theoretically a very large skew of the clock could insert some error during the transfer of the test packet. But, in practice, a discontinuity in the source clock during the time between the taking of the initial and final timestamp is much more likely. This might happen, for example, with certain implementations of the network time protocol (NTP) [126]. The resolution of the clock is always responsible for some uncertainty.
- *Reply delay* at Dst denotes the time required by the Dst from receiving the packet originated from the Src and sending the corresponding response. This systematic error equals the difference in wire-times between the receiving of the first bit of the test packet by Dst and the sending of the first bit of the response by Dst. When the reply delay is known at the Src, the desired metric can be corrected. The NTP, for example, takes into account the round-trip delay between the time server and the host which has to be time-adjusted. The time server needs some time between the receipt of the time request from the host and the sending of the corresponding reply including the current time. This time span is monitored and subtracted from the overall delay in order to correct the round-trip delay by placing the corresponding timestamps in the reply packet.

These errors may occur for every single measurement and are, therefore, observable in the corresponding statistics. In addition to that the overall statistical results may be affected by the threshold (or methodology to distinguish) between a large finite delay and loss.

The calibration and context in which the metric is measured must be carefully considered, and should always be *reported* along with metric results. The items to be considered are the type-P of test packets, the threshold of infinite delay (if any), error calibration, and the path traversed by the test packets. This list is not exhaustive. Any additional information that could be useful in interpreting applications of the metrics should also be mentioned.

The IETF defines a sample as a sequence of single round-trip delay metrics, measured at times taken from a *Poisson* process. This certain process has the advantage of limiting bias, but other methods of sampling might also be appropriate for different situations.

Finally, based on the delay Poisson stream, there are several *statistics* offered, like round-trip delay percentile, median, minimum, or inverse percentile [139].

Beside the IETF the ETSI project TIPHON also gives guidelines on how to determine delay values. Basically, it defines the *round-trip transmission time* [154] very similar to the IETF, but in a more general way:

> "Time in milliseconds for a packet to be transmitted from host A and received at host B and to be retransmitted from host B and received back at host A."

The suggested *statistics* are the minimum and maximum packet transmission times as well as the mean round-trip packet transmission time. Furthermore, TIPHON states that the reflection of a packet for round-trip measurement should be at the protocol layer that the measurement is addressing.

Beside this transport layer measure, TIPHON introduces a more general and user-oriented metric for the perceived speech quality. The so-called *mean one-way delay* is not restricted to IP networks [154]:

> "Mean one-way delay is the time taken in milliseconds for a test signal to go from the near end voice test point, traverse the network, get looped back at the far voice test point and arrive back at the near voice test point divided by two."

The mean delay is derived from at least 10 measurements or from 90 percent of the largest delay measure, whichever is greatest. The delay measurement test signal is illus-

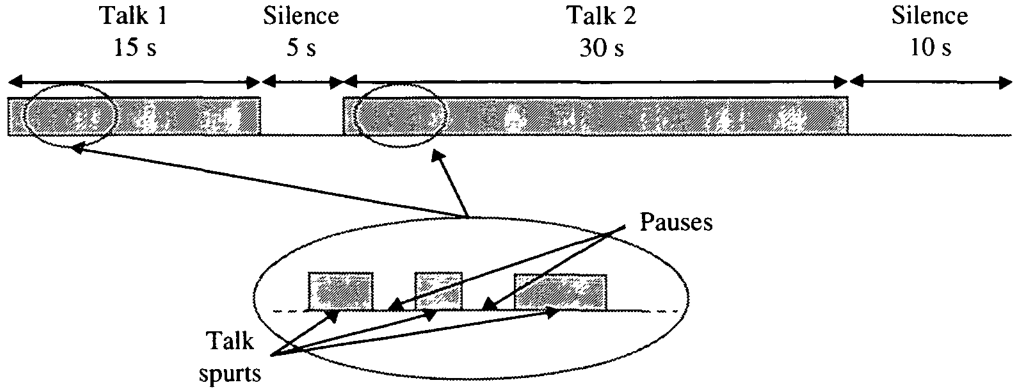

Figure 2.10 – Composition of the TIPHON delay measurement test signal.

trated in Figure 2.10.

It contains periods of speech activity (Talk 1 and Talk 2) and periods of silence. Talk 1 is an initialization sequence, which allows dynamic de-jitter buffers to converge. Both Talk 1 and Talk 2 contain periods of talk spurts and pauses. During the silence intervals the de-jitter buffers are able to adjust their length. The talk spurts consist of either natural speech or an artificial test signal with some speech characteristic. Moreover the average duration of such a spurt should be around 1 s and the average pause about 1.6 s. It is further recommended that the silence intervals in the spurts last for at least 300 ms.

Delay assessments, using cross-correlation or another appropriate technique, are made for each talk spurt. At least 10 measurements are required to determine the TIPHON mean one-way delay measure during Talk 2, which implies 10 opportunities for the de-jitter buffers to adjust. Therefore, this period should be made up of at least 11 talk spurts resulting in an overall measurement time of at least 30 s.

One-Way Delay

The definition within the IETF of the one-way delay through an IP network is very similar to that of the round-trip transfer time [137]:

> "The type-P-one-way-delay from Src to Dst at T is ΔT means that Src sent the first bit of a type-P packet to Dst at wire-time T and that Dst received the last bit of that packet at wire-time T + ΔT."

The corresponding illustration is given in Figure 2.11.

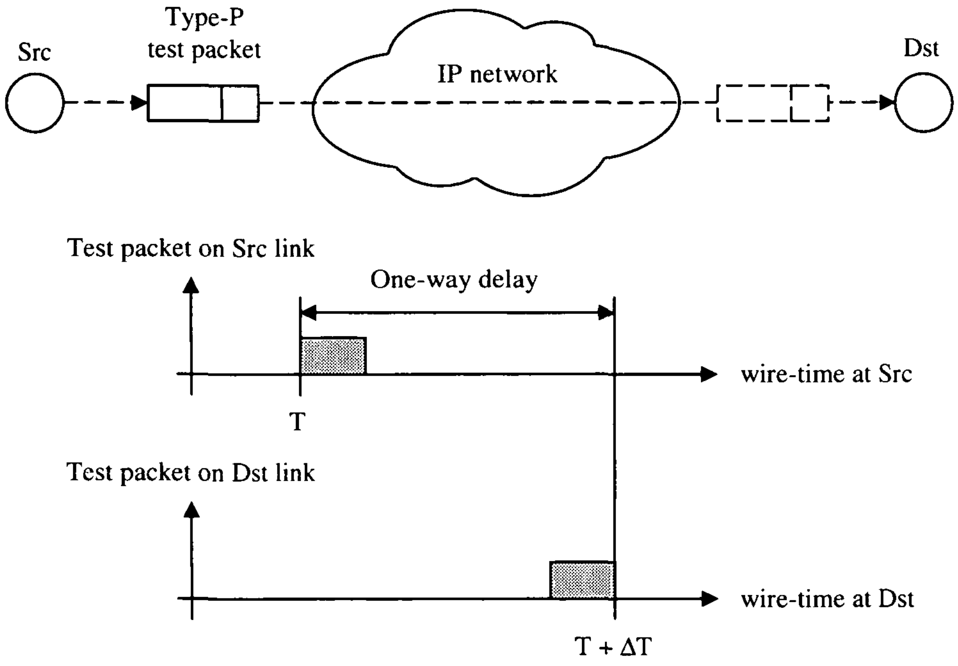

Figure 2.11 – Determination of the one-way delay through an IP network according to IETF. The result is derived from the difference of the sending timestamp at Src and the receiving timestamp at Dst.

The reader is referred to the previous section regarding the meaning of the different terms in the definition, the composition of the test packet, the sources for errors and uncertainties (except the reply delay), the contents of the measurement report, the inter-packet sending interval (Poisson process), and the overall statistics. The estimate of the one-way delay is derived from the difference of the timestamps.

Once again, TIPHON denotes a widely applicable definition of the one-way transmission time [154]:

> "Time in milliseconds between the emission of a signal and the time it is received, includes delays due to equipment processing as well as propagation delay."

Generally, all definitions have in common that Src and Dst host have to be synchronized, i.e. that their clocks are very close to each other. The IETF IPPM working group as well as the ETSI project TIPHON do not specify how to achieve synchronization.

Synchronization

Beside the clock resolution and the difference between wire-time and host-time, synchronization significantly influences the accuracy of the one-way delay measurement. It is defined as follows [137]: "Synchronization measures the extent to which two clocks agree on what time it is. For example, the clock on one host might be 5.4 ms ahead of the clock on a second host." As delay values may also be in the range between 100 µs and 10 ms, a close synchronization of Src and Dst is very important. Basically there are two methodologies to exchange the time information between the hosts, and, furthermore, to obtain the coordinated universal time (UTC) time [39]:

- Global positioning systems (GPSs) achieve synchronization within several 10s of µs. The Information Society Technologies (IST) project "adaptive resource control for QoS using an IP-based layered architecture" (AQUILA) [166], for example, deployed Meinberg GPS cards [167] to guarantee a high level of synchronization.
- Application of the NTP allows synchronization with at least several ms or—in the case of WANs—up to several 10s of ms. The accuracy depends on the stability and symmetry of the delay between the involved NTP agents. Moreover, the paths used for the exchange of NTP messages may also be subject of the measurement. Normally there are several primary NTP servers distributed over the network which obtain the time information from GPS receivers (highest stratum level). With a highly reliable and well-designed network the other NTP servers should yield good results in terms of synchronization by asking the primary servers.

Beside the current time, the clock frequency has to be updated in order to correct the so-called clock skew (also defined as clock drift), which determines the change of accuracy (or the change of synchronization) with time. Accuracy defines the extent to which a given clock agrees with UTC [137]. In computers, the frequency of a crystal oscillator depends on its shape, size, temperature, etc. Therefore, its drift rate remains mostly constant when surrounding conditions such as temperature are stable. Typical drift rates of crystal oscillators compared to UTC are in the order of 100 µs per day.

GPS allows correcting the system time continuously. Thus, an erroneous clock frequency can be neglected in contrast to an NTP synchronized system. Further details on synchronization are found in Section 0.

2.2.4 Measurement Devices

Figure 2.12 points out an example for localized one-way delay measurement through a PSTN network with an IP backbone between two gateways using the Agilent voice quality tester (VQT) [168].

The Agilent VQT is not able to perform a distributed one-way delay measurement, as both endpoints of the connection have to be within the range of the device. That's why the measurement does not rely on synchronized end-points, which, consequently, increases the achievable accuracy. The VQT injects an artificial pseudo-random, pattern-repeating noise, and records the system output at the same time. The transmit signal covers a broad spectrum of frequencies within the telephony band. Moreover, its auto-correlation function comes with a distinct peak, which enables a reliable determination of the one-way delay. Both signals, the transmit and the receive sequence, are taken to perform a signal cross-correlation (see Figure 2.12). The VQT determines the highest concentration of energy within the distribution of the correlation, and identifies the corresponding time value ΔT as

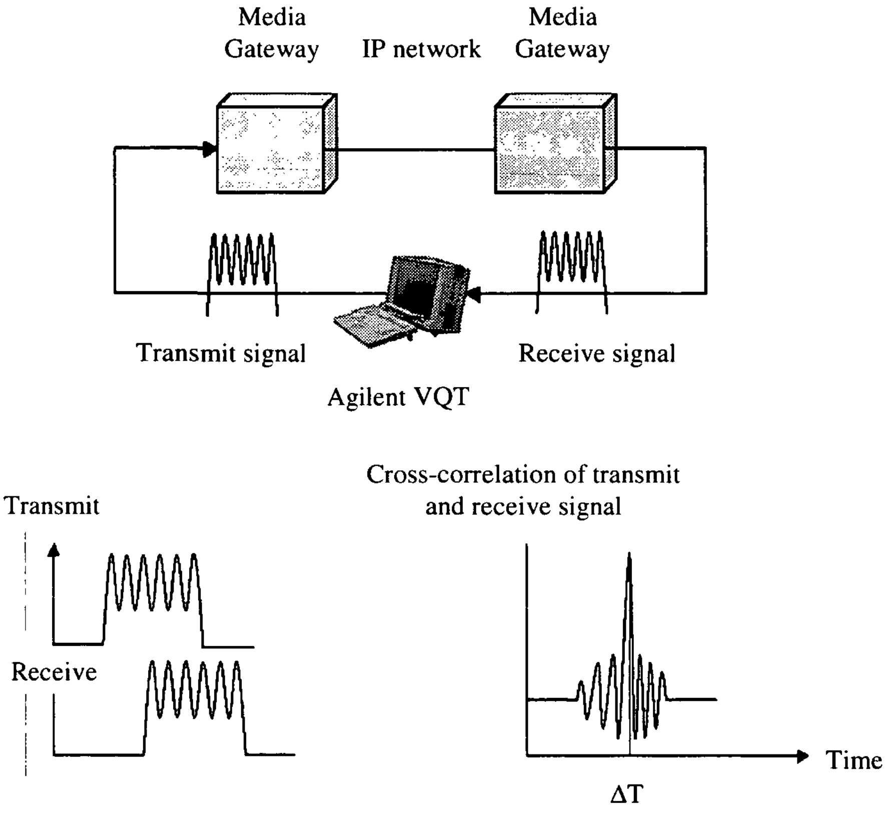

Figure 2.12 –The Agilent VQT cross-correlates the transmit and receive signal in order to determine the one-way delay.

Device	VQT [168]	NetQual [169]	OPERA [170]
Vendor	Agilent	SwissQual	Opticom
Round-trip Delay	X	X	-
One-way Delay	X	–	X
Accuracy	± 1 ms	± 10 ms	± 10 ms (NTP) ± 3 ms (DDLC)

Table 2.5 – Measurement devices, their vendors, the kind of assessable delay, and the achievable accuracy.

the overall delay of the signal. The advantages over other methodologies that measure delay such as an acoustic ping or tone are: A more robust technique of time synchronization, less sensitivity to front-end clipping, noise, attenuation, and loss as well as a measurement across all telephone-band frequencies.

Beside the VQT, the NetQual performs only round-trip delay measurements with a loop-back function at a second NetQual device in order to reflect the test signal. The objective perceptual analyzer (OPERA) device is able to carry out distributed one-way delay measurement, whereby the synchronization is carried out by deploying NTP. Moreover, OPERA has implemented the dynamic driver latency compensation (DDLC) algorithm, which improves the uncertainties due to operating systems inefficiencies down to ± 3 ms for 99.5 percent of all results.

2.3 Echo Cancellation

> The term echo represents "a wave that has been reflected or otherwise returned with sufficient magnitude and delay to be perceived".[17]
>
> Echo cancellation denotes "the reduction of the power level of an echo or the elimination of an echo".[17]

2.3.1 Echo Terminology

Echoes have a strong influence on the conversational voice quality (see Section 2.1.5). The term *echo* can be unambiguously differentiated from the term *reverberation*, as echo defines "a discrete repetition of a sound, as opposed to reverberation, which is a continuous

---

[17] Glossary of Telecommunication Terms-Federal Standard 1037C

wash of closely spaced, non-discrete echoing sound"[18]. The more precisely named term *discrete* or *distinct echo* identifies a speech wave that arrives only once a distinct time gap (e.g., a few tens of milliseconds) after the direct sound. Beside the delay from the talker's mouth to the talker's ear the degree of annoyance also depends on the echo level (see Section 2.3.3). Reverberated sounds always come with some amount of spectral distortion. Most people prefer some amount of reverberation to a completely anechoic environment, but this depends on the application. Another meaning of a reverberated sound is expressed by: "continuation of a sound after the sound source has stopped due to reflection of sound waves within a closed area"[19].

If the propagation delay of the sound wave in telecommunication systems decreases down to about 10 ms or less, it will be heard as sidetone, i.e. the talker hears his or her own voice instantaneously in the earpiece of the telephone. The absence of this side-effect would be disturbing, because people think that if they can't hear themselves, the person at the other end can't hear them either. In the interval between approximately 10 ms and 30 ms the speech undergoes a hollow or "tunnel" sound characteristic [15].

If the talker also perceives a reflected, delayed and possibly spectrally distorted version of the original wave, he or she is faced with *talker echo*, which is the most important form of reflection regarding telecommunication systems. It is also possible that the listener hears the talker's voice twice—a loud signal first, then attenuated and much delayed. This is identified as *listener echo* (also see Figure 2.13).

Furthermore, the reflection of interest can be caused by electric or acoustic coupling. Electric echoes are originated by impedance mismatch in interfaces between the two and four-wire portion of a network.

Acoustic Echo

One has to cope with acoustic echo, if some amount of the acoustic signal of the loudspeaker is coupled into the microphone of the same device. This parasitic signal lies about 10–15 dB below the signal of the person who talks into the microphone [8]. In the case of an IP phone, an IP client, or an ISDN device there are not any hybrids involved. Thus, the main sources of echoes in such fully-digital connections are acoustical. The corresponding terminals are characterized by the terminal coupling loss (TCL) or by the weighted TCL (TCLw), whereby, for example, most ISDN phones have a TCL value of about 45 dB. The parameters TCL and TCLw are measured with respect to the input and output of a digital telephone set and they include the following coupling effects within the telephone sets [148]:

- Primarily the acoustic coupling at the user's interface has to be considered. In this context there are two modes of operation for telephone devices. In the handset mode the acoustic path between receiver and transmitter capsule becomes important, while in the case of a hands-free situation acoustic coupling occurs between loudspeaker and hands-free microphone.
- Crosstalk via capacitive coupling within the handset cord.

---

[18] Babylon Glossary of Electronic Music Terms

[19] Babylon German-English dictionary

- Mechanical coupling (structure-borne sound) due to the mechanical parts of the terminal.
- Coupling through the power supply for codec and amplifiers.

Basically an analog telephone set, which is connected to the local switch via the two-wire customer loop, shows the same effects with the exception of the coupling caused by the power supply. Because of the high gap between the loudspeaker and microphone in headsets and the almost missing mechanical coupling, acoustic echoes don't arise with the use of those devices. With an appropriate echo control device it is feasible to reduce the power of the acoustic echo well below the speaker's signal, which corresponds to an attenuation of 45 dB or more, even when deploying a hands-free phone. Single-channel acoustic echo cancellers are already in wide-spread use, but the more difficult problem of multi-channel (e.g., stereo) acoustic echo cancellation is still subject of research activities.

Echo cancellers are designed either to compensate for electric or to cope with acoustic echo, but not both kinds of reflections. The reason is that the acoustic echo-path[20] varies much more than the electric one because of [1]:

- Non-linearities in the echo-path, i.e. the loudspeaker-room-microphone system.
- The much longer and time-variable impulse response of an acoustic echo-path.
- The higher influence of background noise.

Typical values for echo-paths in terms of reverberation time and volume of the test room are given in the corresponding ITU-T Recommendation [79] with the intention to provide a common basis for different test procedures.

Electric Echo

Mismatched impedances between the two- and four-wire part of the PSTN are responsible for electric echoes (also known as hybrid, line or network echoes). In local telephone calls electric echoes do not degrade the voice quality because the echo levels are small, and the delays of the talker echoes are below 10 ms. In a connection over longer distances in which the end-to-end delay is non-negligible (i.e., 30 ms or more), distinct echoes are perceived by the user. The reflection in such circuits takes place—as already mentioned—at a device called hybrid. It is placed in the local switch and connects the two-wire twisted-pair copper lines line of analog sets to the four-wire portion of the circuit-switched network. Such a telephone connection is illustrated in a highly simplified manner in Figure 2.13 [42]. Usually the backbone of the PSTN uses digitized voice samples and is, thus, seen as virtual four-wire. The hybrid is discussed in depth in Section 2.3.2.

The four-wire path of the network offers potential feedback loops for the reflection of echoes at the hybrids. That's why the hybrids should insert a sufficient loss in this path to avoid a degradation of the voice quality or under exceptional circumstances even oscillation. The latter effect is also called singing and it starts, when the sum of all losses and gains is equal or less than 0 dB [146]. Such effects are only observed in mixed analog and digital connections—as widely implemented in the PSTN all over the world—where gains

---

[20] The term "echo-path" instead of "echo path" is used, when referring to the path the acoustic or network echo control device is working on.

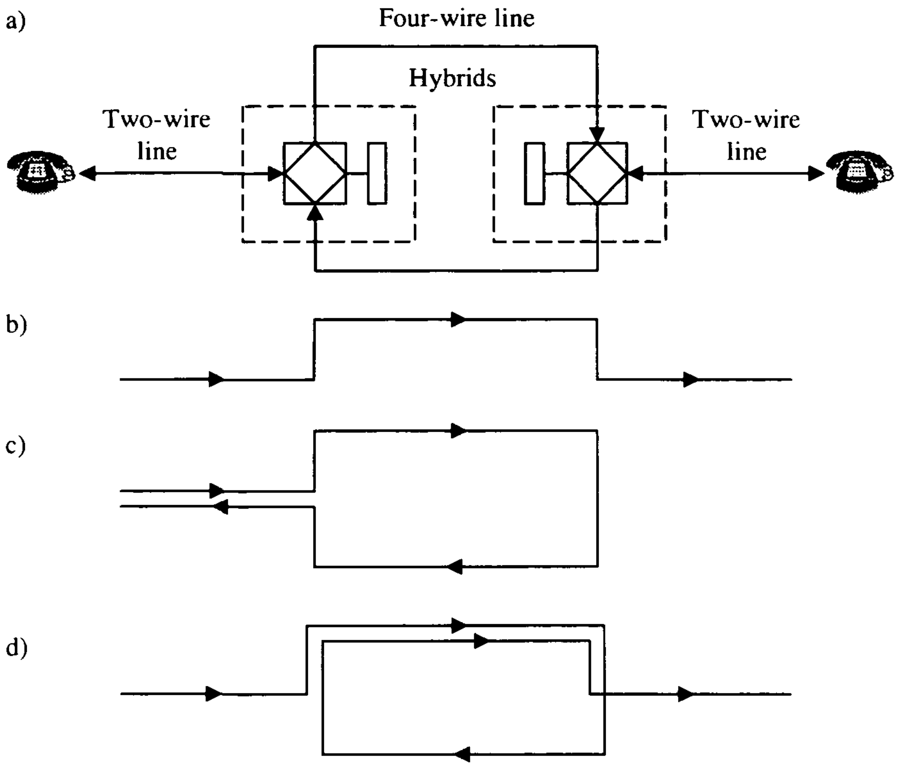

Figure 2.13 – Sources of echo in the telephone network. (a) Simplified telephone connection. (b) Talker speech and (c) talker echo path of the user on the left-hand side. (d) Listener echo of the other participant.

are inserted in the digital four-wire path. Further explanations concerning the stability criteria and the broadened frequency spectrum of this oscillation are written in the corresponding ETSI Guide [146]. Beside the large loss along the loop, the hybrid should not affect the two talker speech paths to a noticeable extent. Figure 2.13 (b) illustrates the speech path of the talker located at the left end of the connection. The two generally stated distinct echo mechanisms above—the talker and listener echo, respectively—are shown for this typical telecommunication system in Figure 2.13 (c) and (d).

2.3.2 Hybrid

In many countries, four-wire links are used in the transit network between the local switches. The *hybrid* (also known as duplexer [8], induction coil, echo-path [80], two to four-wire transducer, hybrid or differential transformer [47]) represents the interface between the two-wire and four-wire portion of the network. Different impulse responses of this device and the corresponding values are given in [80]. They are mainly characterized by the echo delay, the overall attenuation, and the length of the impulse response. The common way to symbolize this device is shown in Figure 2.13, where each corner of the square represents two wires. The separation into two- and four wire portions of the network takes place at the local switch, where analog phones are connected to the central office (or local switch) by the two-wire line (also called the customer, subscriber or local loop). Echoes are generated, when some signal on the receiving path on the four-wire side of the hybrid is coupled into the sending path. It is possible to communicate over a two-wire line in both directions at the same time, while a considerable amount of wiring as well as of local switching equipment can be saved. A local call is established by connecting two local loops at the central office equipment. At distances above 50 kilometers, it is necessary, to amplify the transmitted signals [1]. As amplifiers are one-way devices, separate paths are needed for both directions. Another reason for the four-wire backbone in the PSTN is that calls can be transported more efficiently when they are time-multiplexed. Multiplexing over wide-band transmission channels requires different time slots for the sending and receiving signals.

Beside the deployment in the local switch, a hybrid also transforms the four wires of the loudspeaker and the microphone within an analog telephone set into one two-wire pair. There is a small imbalance kept in the phone to produce a sidetone. Values for the corresponding sidetone masking rating (STMR) can be found in [150]. Consequently, the simplified telephone system of Figure 2.13 can be detailed in terms of sources of talker echoes as shown in Figure 2.14 [58].

The electric echoes generated by the hybrids located in the local telephone A and in the local exchange A have only small delay (well below 10 ms). Therefore, both are perceived as sidetones. The reflections formed by the transformers in the far exchange B and in the far telephone B as well as the acoustic and other possible echoes caused by the far telephone set may disturb the local speaker, when they are coming after 30 ms or later (without echo cancellation) [8] [15].

Basically the hybrids in Figure 2.13 and Figure 2.14 are based on a bridge circuit; the

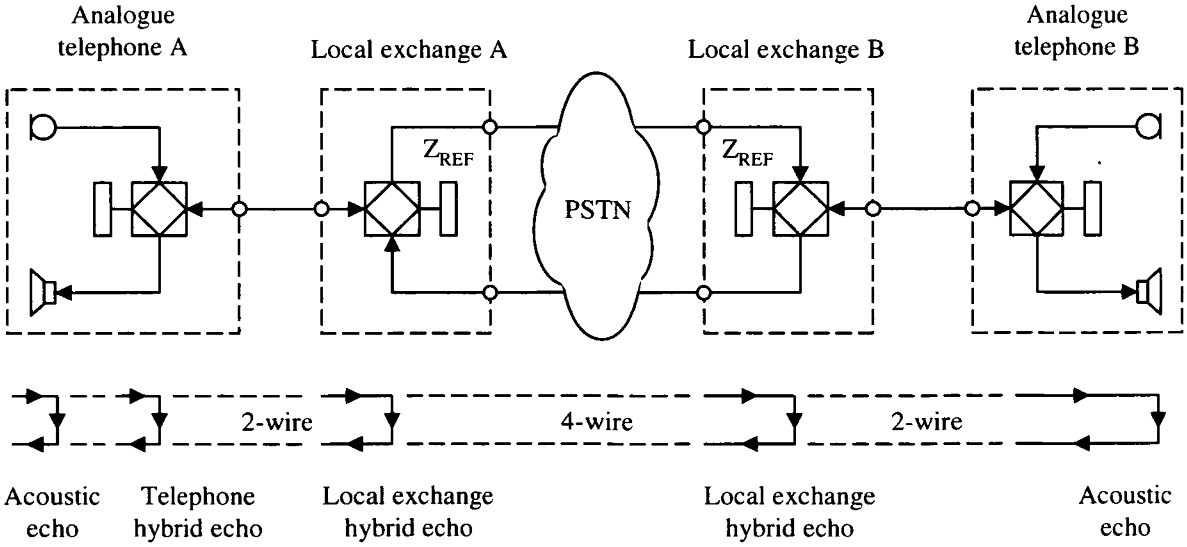

Figure 2.14 – Talker echoes in the PSTN are originated at hybrids and due to acoustic couplings in the telephone sets.

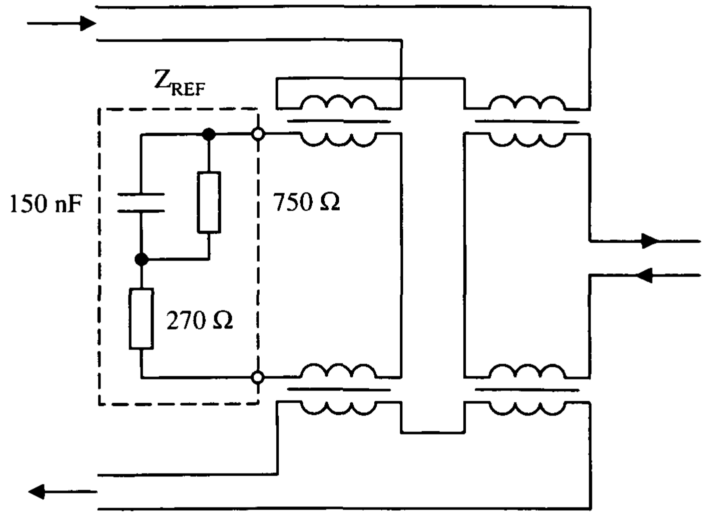

Figure 2.15 – Detail of a hybrid.

hybrid placed at the right-hand or listener's side in both figures are detailed in Figure 2.15 [47][9]. An integrated version of this bridge is shown in [8]. Ideally the hybrid contains two identical transformers and the reference impedance Z_{ref} (also called balancing impedance), which matches the characteristic impedance of the two-wire line at all frequencies. Under such ideal conditions all signals on the ingress-path of the trunk side are completely coupled to the two-wire circuit and generate no echoes on the egress-side of the four-wire path. ETSI defined the "European harmonized complex impedance" as reference Z_{ref}, which is made up of 270 Ω in series with a parallel combination of 750 Ω and 150 nF [151].

Signals originated in the two-wire circuit are transmitted to both sides of the four-wire path. But they have no effect on the incoming circuit, as they face the output of an amplifier. If there is a mismatch between the reference and the two-wire circuit impedance, parts of the incoming signal are coupled to the egress of the hybrid, thus resulting in an echo. The reflected signal is not just attenuated, but also a filtered version of the input signal. The impedance of the local loop is not constant as it depends on various parameters like the length, type of wire, number of phones attached to it, etc. Assuming statistically distributed call requests the telephone network is always designed with far fewer four-wire circuits than customer loops. The assignment between both parts of the system occurs dynamically and so some impedance mismatch always exists.

2.3.3 Echo Tolerance Curves

The main signal parameters that determine the degree of annoyance of talker echo are the delay experienced by the echo from the mouth to the ear of the speaker and the level difference between the talker's voice and the received echo signal. This level difference is called the talker echo loudness rating (TELR) as described in [69] and was former known as overall loudness rating (OLR) [73].

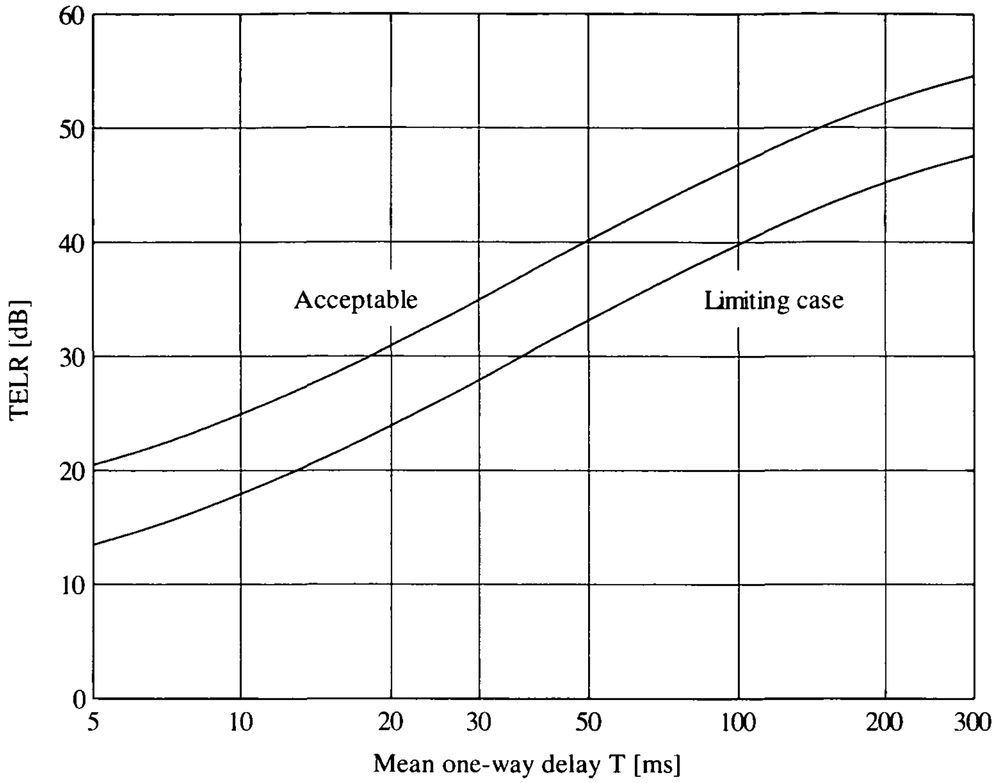

Figure 2.16 – Talker echo tolerance curves. The required overall echo attenuation TELR increases with higher mean one-way delays T. The "acceptable" and "limiting case" curves are displaced by 6 dB.

The American Telephone and Telegraph Corporation (AT&T) finished an extensive study on echo tolerance curves in 1971, which has been published first in [74]. The corresponding diagrams show the required TELR values over the mean one-way transmission time T for a constant percentage of disturbed talkers. The one percent and ten percent graph have been renamed as "acceptable" and "limiting case" curve, respectively; both are shown in Figure 2.16 (see ITU-T Recommendation G.131 [75]). One percent, for example, means that on average one percent of the users complain about objectionable echo. System design should follow the "acceptable" graph, while the "limiting case" should only be applied in exceptional circumstances. Furthermore, the figure clearly shows that echo becomes more audible as delay increases. Therefore, all telephony technologies that insert high delays have to cope with echo problems. Most compressed voice technologies, satellite transmissions, and, IP telephony are coming with noticeable delays. IP based telephony connections face the same echo level and delay trade as any other telephone system.
The curves in Figure 2.16 can be approximated by the following equation [148]:

$$\mathrm{TELR}(T) = A_0 - 6e^{-0.3T^2} + 40\lg\frac{1+\frac{T}{10}}{1+\frac{T}{150}}, \tag{2.7}$$

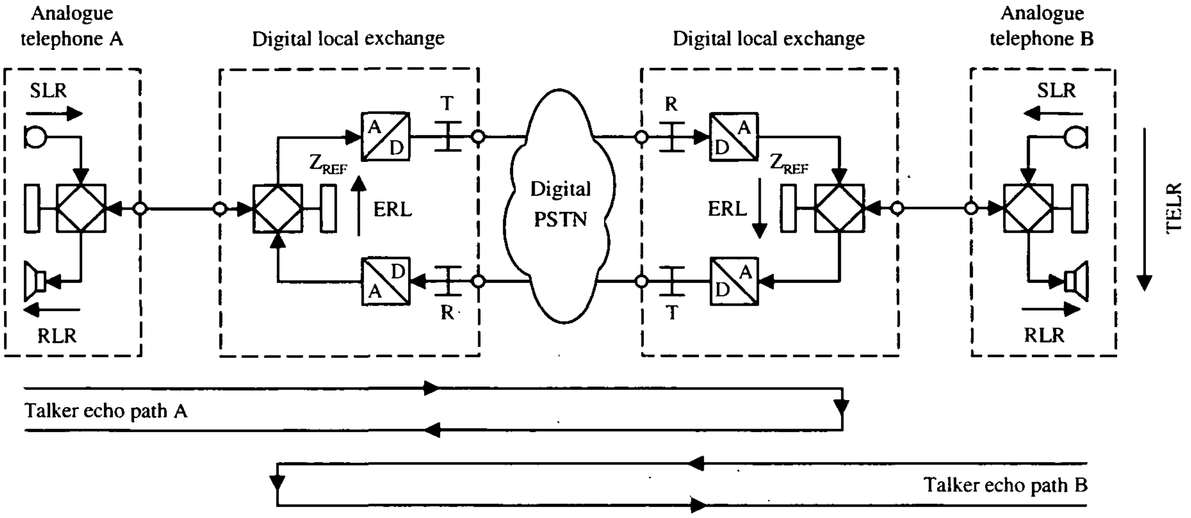

Figure 2.17 – Talker echo path and losses for an analog telephone system connected to a digital backbone through local switches.

where T denotes the mean one-way transmission time in milliseconds, and A_0 = 14 dB is used for the “acceptable” graph and A_0 = 8 dB for the “limiting case” curve. The equation is also applicable for delay values below 10 ms. However, in the case of T < 1 ms, the talker echo is perceived as sidetone, thus permitting much higher values for the talker echo (indicated by the STMR). The consequences of talker echo on the perceived voice quality are also predicted in the E-model; an example for different TELR values over the one-way delay is given in [157].

The talker echo path of a widely deployed mixed analog and digital connection is given in Figure 2.17.The sending loudness rating (SLR) and receiving loudness rating (RLR) model the acoustic to electric efficiency of the receiver and the emitter respectively. Louder telephones, for example, show lower SLR and RLR values. Values for European and North-American countries are given in [150] and [156], respectively. In the case of software IP phones these values are affected by the sound card settings. The echo return loss (ERL, also known as trans-hybrid loss [148], echo balance return loss, or return loss [73]) has been introduced in [80] and represents the amount of echo that is reflected by the hybrid. The parameter ERL is determined by the loss between the input and the output of the four-wire side of the hybrid. Corresponding values must not be smaller than 6 dB; typically they are in the range between 10 dB and 20 dB [1].

The TELR of talker echo path A is derived from the sum of attenuations of the different components placed along the echo path from the talker’s mouth to the talker’s ear by using analogue telephone A:

$$\mathrm{TELR} = \mathrm{SLR} + 2\mathrm{T} + \mathrm{ERL} + 2\mathrm{R} + \mathrm{RLR}\,. \tag{2.8}$$

The loss plan for a fully analog or a mixed analog/digital connection introduces the transmit (T) and receive (R) loss[21]; both are placed in the network with the intention to have a 0 dBr point at the exchange. Many European countries use T = 0 dB and R = 6 dB or R = 7 dB [72]. Moreover, the additional losses T and R are a balance between enough loss to attenuate the echo and, thus, maintain circuit stability, while still providing an adequate signal level over a range of analog loops [157].

In the case of an all-digital connection (e.g., with an ISDN phone) the TCL of that device is used instead of the ERL of the hybrid and no additional losses are introduced in the network:

$$\mathrm{TELR} = \mathrm{SLR} + \mathrm{TCL} + \mathrm{RLR}\,, \tag{2.9}$$

where most digital handsets have a TCL typically above 35 dB, and often in the 40–46 dB range [8].

Echo Delay

ITU-T recommends that echo control devices shall be installed on all connections which exceed one-way echo delays of 25 ms [75]. However, in networks with round-trip delays of at least 30 ms and with relevant echo levels, the deployment might already improve the conversational voice quality [15]. Delays in telecommunication networks are a consequence of (see Section 2.2):

- Long distances: On international circuits, in large countries like the US, or on connections over satellite links echo control is required.
- Packet, frame, or cell based systems such as IP, frame relay, or ATM networks are coming with relevant delays.
- Speech compression: Various coding schemes need time to remove more or less redundant information from the voice stream. Speech coding with low algorithmic and processing delay relaxes the requirements for the performances of the echo control device.

Basically, there are two techniques used for controlling network and acoustic echo—echo suppression and echo cancellation. The first method was developed over 70 years ago and has been widely replaced by the latter one, which was invented in the 1980s.

2.3.4 Echo Suppressor

This echo control device, which has been standardized in [77], worked well on communication circuits with round-trip times of less than about 100 ms [1]. This value corresponds to distances of a few thousand kilometers. Nowadays echo suppressors are found in completely analog networks and in low-end, hands-free phones [8]. They are relative simple voice operated switches and function as shown in Figure 2.18.

A large loss is placed in the transmit path of the talker situated at the so-called near end (also defined as local end, drop side [8], or local tail), when the control unit detects a significant amount of speech on the receiving path (compared to the sending path). The level

---

[21] The additional losses are not explicitly mentioned in the rest of this thesis, as they are regarded as part of the ERL; this assumption also corresponds to the ITU-T Recommendation for network echo cancellers [80].

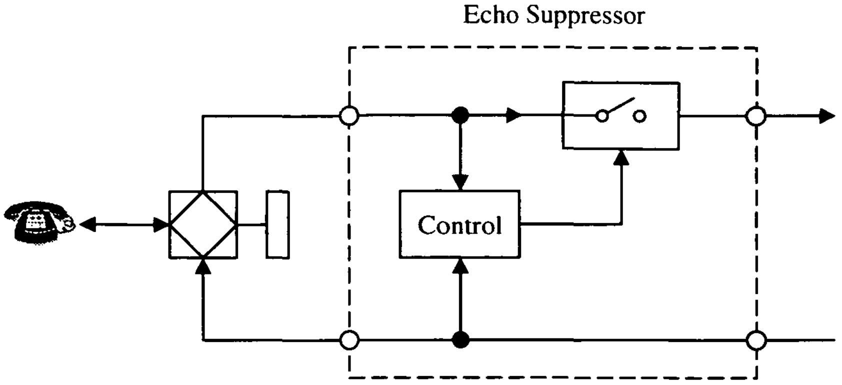

Figure 2.18 – Echo suppressor.

of the returned echo originated by the talker at the far end (also called distant or remote end) is almost completely suppressed in this way. But, at the same time, the signal originated at the near end (often denoted as double talk signal) is prevented from reaching the far end. Hence the original design of echo suppressors was not able to transmit speech signals from both talkers simultaneously; only half-duplex mode of communication and no double talk operation is feasible. One way to accomplish double talking is to inhibit echo suppression in these special talking periods [1] so that far end participant perceives the disturbing talker echo superposed to the speech from the near end talker. But the full level of talker echo may disturb the communication significantly. Another possibility is to decrease the sending attenuation to a smaller value and, additionally, to insert a small amount of attenuation in the receiving path [15]. This enhancement of the basic concept (not illustrated in Figure 2.18) reduces sufficiently the echo level through both losses, while attenuating the sending signal only slightly.

The design of the control mechanism critically influences the performance of the echo suppressor. This unit has to decide, which talker is dominant at one time in order to inhibit the transmission of the other talker's signals. To accomplish this function, it has to distinguish between echo components caused by the far end talker and speech signals from the near end on the sending path. Both signals are speech signals with similar statistical properties. The only attribute for decision is the level of the signals. A wrong detection may lead to high talker echo levels or considerable speech from the local end might be switched off temporarily. The latter effect is well known as *clipping*, whereas front-end clipping denotes an incomplete transmission at the beginning of a voice stream and back-end clipping occurs, when parts at the end of the last word are suppressed. When utterances in the middle of the continuous voice stream get lost, one has to deal with speech gaps. Because of those voice degrading effects, echo suppressors have been superseded to a large extent by the superior echo canceller[22].

---

[22] Instead of "echo canceller" as used within the ITU-T another spelling often found in literature is „echo canceler“ [1].

2.3.5 Echo Canceller

Another motivation for the introduction of this echo control device arose with the advent of commercial communications satellites in 1965. The telephone signals about a geostationary satellite link experience a round-trip delay of 500–600 ms. With such long delays the way of conversation changes, which leads to an increasing number of errors caused by the echo suppressor. As very large scale integration (VLSI) implementation became available at a reasonable price in the 1980s, the widespread deployment of echo cancellers started. As an alternative to satellite links, the introduction of fiber optics avoided long delays. Nevertheless, in the last decade, because of the increasing number of speech coding in telecommunication systems, delay has again become an issue. Other modern trends such as packet or cell based transmission systems as well as wireless communications are also inserting relevant delays. Therefore, echo control evolved to a topic many system designers in the field of telecommunication are confronted with.

Echo cancellers as well as echo suppressors for different applications are placed as close as possible to the source of echo. Furthermore, they have in common that—since they remove echo from their local tail—the delay over the transmission medium (e.g., the network or air interface in mobile communication) has no influence on their performance.

Basically there are two possible applications of echo cancellers; the compensation of acoustic and the reduction of electric or network echoes. The corresponding devices are named acoustic and network echo canceller.

Acoustic Echo Canceller

As already explained in Section 2.3.1, the acoustic echo-path varies much more and is highly non-linear, which makes the dynamic adaptation of the loudspeaker-room-microphone system more difficult. The main difference to an electrically coupled path is the usually much longer impulse response (typically 50–300 ms [1]) and that it may change its behavior rapidly at any time (e.g., because of a moving person or an opening window). Thus an acoustic echo canceller [79] requires a longer transversal filter to cancel the echo and adaptive filters with higher convergence speeds have to be implemented. Both reasons are responsible for the higher computation power—compared to network echo cancellers—to fulfill the requirements, which are met only partly by the current products and prototypes. Acoustic echo cancellers are built on the same basic concept as network echo cancellers (see also Figure 2.19). In order to increase the convergence speed of the adaptive filter acoustic echo cancellers are equipped with pre-whitening filter and a subband concept is also applied [1].

Implementations of acoustic echo control devices are found, for example, in hands-free terminals and in modern mobile devices. The transmission characteristics, speech quality parameters, and a categorization of the hands-free terminals are determined in [99].

Network Echo Canceller

The new approach dealt within this thesis is based on digital network echo cancellers (also known as line or electric echo cancellers), which are defined by ITU-T Recommendation

G.168 [80] as follows:[23]

> "A voice operated device placed in the 4-wire portion of a circuit and used for reducing the cancelled end echo present on the sending path by subtracting an estimation of that echo from the cancelled end echo."

It is integrated in communication networks with the intention to eliminate echoes reflected at hybrids. In addition to the different expressions for the near end introduced for echo suppressors, the term "cancelled end" is exclusively used for echo cancellers and denotes "the side of an echo canceller which contains the echo-path on which this echo canceller is intended to operate" [80]. Equally, the far end is specified as the non-cancelled end.

Figure 2.19 shows the network echo canceller deployed at one end of the connection, another echo canceller is applied at the other end:

Network echo cancellers as well as echo suppressors are designed to compensate electric echoes and both are placed in the four-wire path of the network. The echo canceller is much more complex in terms of the computational requirements, because of the complex filter algorithm needed for the adaptation process. The adaptive filter has to approximate the impulse-response of the hybrid continuously by updating a linear model of that echo-path, because the corresponding electric circuit between the egress port of the receiving path R_{OUT} and the input port of the sending path S_{IN} is time-variant.

The error signal at the output of the subtractor is only determined and minimized by the filter algorithm, when the distant party is active (i.e., in single talk situations with speech signals originated at the far end). The double talk detector (DTD) accomplishes this task by continually monitoring the incoming signals of the echo canceller as well as the signals at the ingress port of the sending path. Based on the observed levels the DTD decides, whether there is a double talk or not. When detecting a double talk period, the DTD imme-

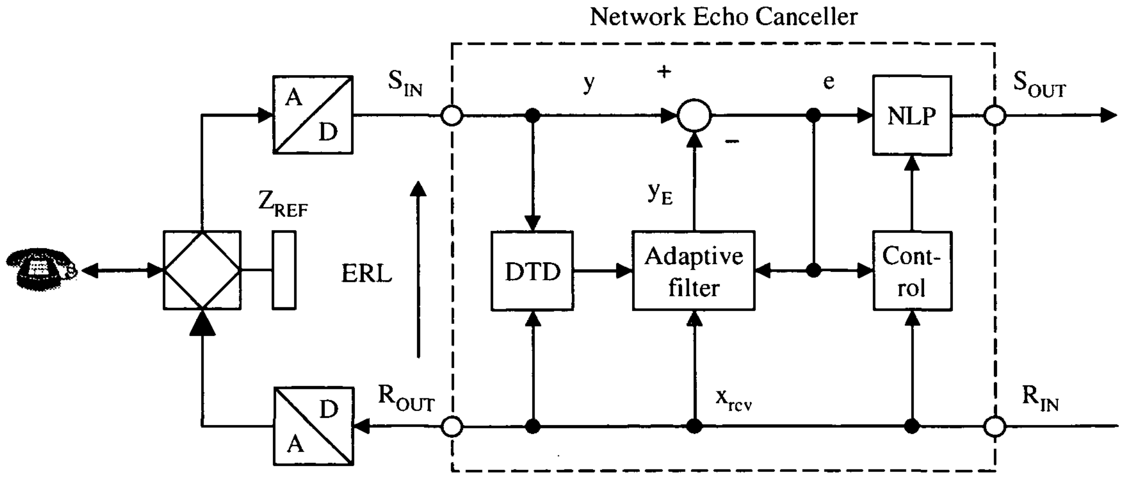

Figure 2.19 – Network echo cancellers continually subtract echo estimates y_E, which are derived from a model of the two- to four-wire hybrid, from the signal from the near end y.

---

[23] In 1997, the ITU-T Recommendation G.165 [78] has been superseded by G.168.

immediately stops the adaptation of the filter (i.e., the step size of the filter algorithm is set to zero, and, in this way, the filter coefficients are held in the current state) in order to prevent the filter coefficients from diverging. Furthermore, the non-linear processor (NLP) is switched into transparent mode in the case of double talk, i.e. all signals are passing through unchanged. The filter carries on with the adaptation process, when there is single talk from the non-cancelled end again. As it has to distinguish between signals from the local end and signals reflected at the hybrid, the DTD is faced with similar problems as the control block of the echo suppressor in Figure 2.18.

The maximum echo attenuation provided by the adaptive algorithms is not high enough to cancel the echo below the perception threshold. A supplementary signal processing unit—the NLP (see below in this section)—introduces an additional attenuation. Current specifications for network echo cancellers [20] [80] recommend residual echo levels below -65 dB_{m0} for speech levels between -30 dB_{m0} and -10 dB_{m0} (with NLP enabled).
Echo cancellers are characterized by the following performance parameters [4]:

- The cancellation depth denotes the achieved overall reduction of the echo level. It is measured in dB and divided into the attenuation provided by the adaptive filter, which is called echo return loss enhancement (ERLE in dB), and the amount of echo reduction by the non-linear processor. As the NLP operates non-linearly, the attenuation in dB can only be calculated in an average sense.
- The window size or echo-path capacity indicates the maximum length of the impulse response of the echo-path that can be modeled by the filter.
- The convergence time specifies the time it takes the echo canceller to adapt to the echo-path with a certain level of echo reduction. The power density spectrum of the signal at the ingress mainly determines the convergence speed of echo cancellers [80].
- The double talk robustness is observed in the case of simultaneous talking on both ends of the connection, where the echo canceller has still to grant for adequate attenuation.

Moreover, echo cancellers have the following main advantages over echo suppressors:

- Better transparency of the sending path.
- Less impairment due to NLP hangover time.
- No loss in the receiving path of the echo canceller.
- The cancellation goes on during talking.
- Well designed echo cancellers can be cascaded.

Applications

This section presents some further examples in which acoustic or network echo cancellers are deployed. Generally, echo control is required, when a considerable amount of delay is experienced by the reflected voice stream (i.e., one-way delay above 25 ms) and when electric or acoustic interfaces induce objectionable echo levels. Telecommunication systems with relevant delays have already been mentioned above.

Network echo cancellers, for example, occur in an international switching center for a satellite link or in the MSC for digital cellular applications. In the global system for mobile communication (GSM) phone system, the voice needs about 100 ms to travel from the

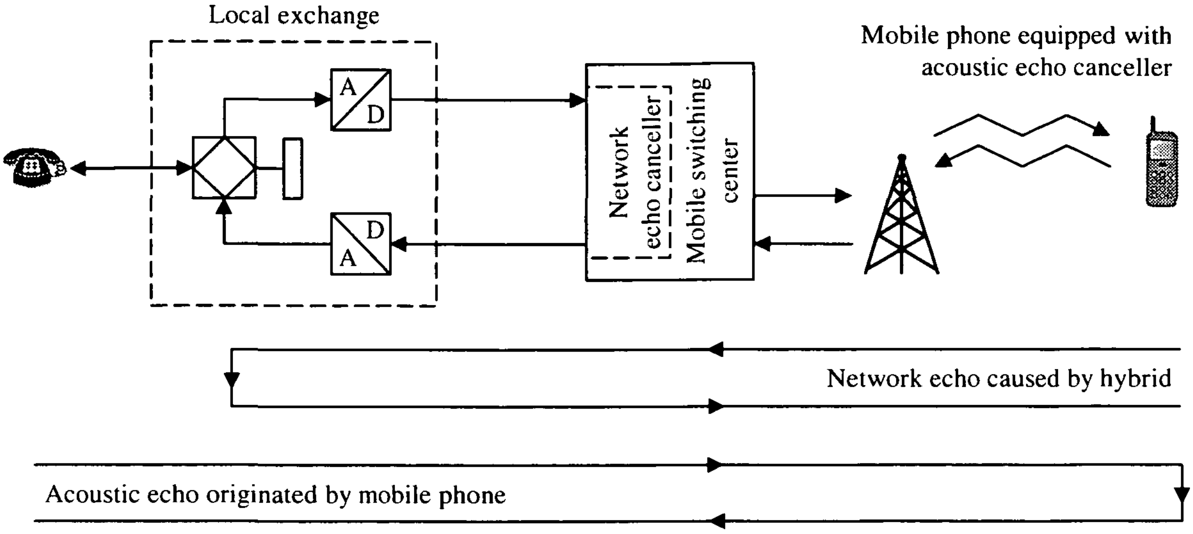
Local exchange
A
D
D
A
Network echo canceller
Mobile switching center
Mobile phone equipped with acoustic echo canceller
Network echo caused by hybrid
Acoustic echo originated by mobile phone

Figure 2.20 – Network echo cancellers are mandatory in the mobile switching center of a digital cellular system and acoustic echo cancellers are deployed in the handsets because of the coupling effect between loudspeaker and microphone.

talker's mouth to the listener's ear, due to [8]:

- The frame length of 20 ms.
- The speech processing (coding, ..) in the mobile device introduces delays of 20 ms, which may be shortened by a powerful DSP.
- Interleaving for channel protection.
- Buffering and decoding.

Because of the short distances in small countries like Austria (resulting in one-way delays below 10 ms) echo control is not necessary for national calls. Thus local exchanges—in contrast to the switches for international calls—are not equipped with echo cancellers. Therefore, the mobile switching center has to provide for adequate echo cancellation (see Figure 2.20) in order to avoid unpleasant talker echoes for the mobile subscriber. These reflections are originated at the hybrid located in the switch of the analog wire line party. Furthermore, acoustic echoes arise especially in modern mobile appliances because of the very small and compact design, which is responsible for the strong coupling between loudspeaker and microphone (structure-borne sound). That's why they contain an acoustic echo canceller; otherwise the PSTN user would be faced with annoying acoustic echoes.

In the case of a VoIP system connected to the PSTN, the electric echo has to be canceled in the IP telephony gateway in a similar way as before (Figure 2.21). The telecommunication takes place between the user of the plain old telephone service (POTS) and a party using a multimedia PC. The latter introduces a significant amount of acoustic echoes, when he or she uses a hands-free terminal consisting of microphone and loudspeaker. Beside the network echo canceller deployed in the gateway, consequently, an acoustic echo canceller is required in the PC environment. To cope with the acoustic echo, generally, an acoustic echo canceller must be installed in every hands-free terminal whatever network

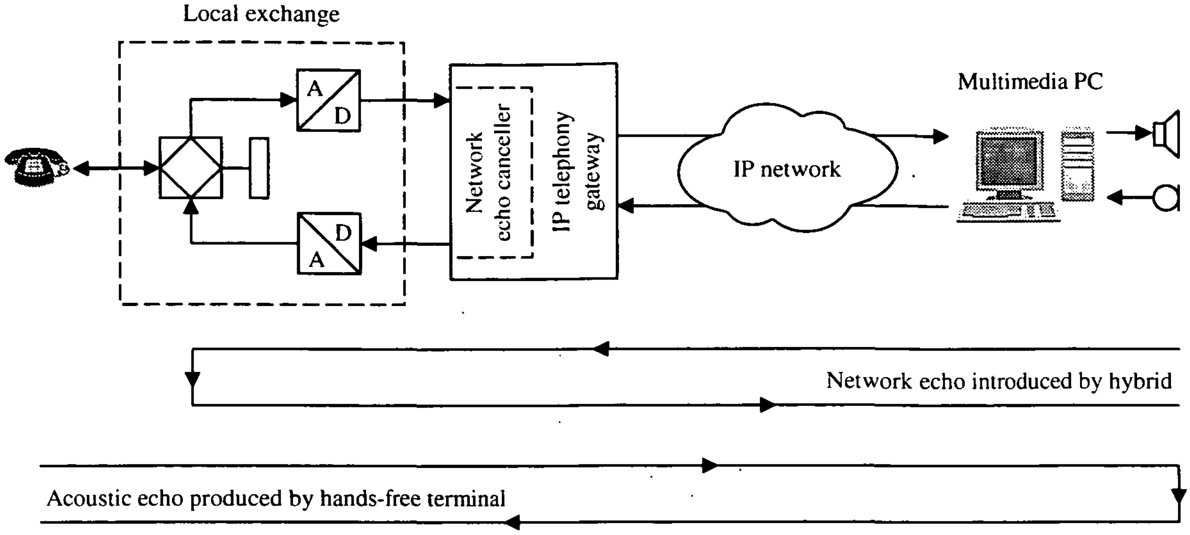

Figure 2.21 – VoIP scenario with an acoustic and a network echo canceller deployed in the hands-free terminal and in the IP telephony gateway, respectively.

and transmission link are used. In this context a critical mobile application is, for example, a hands-free set deployed in a car environment.

Filter Algorithm

The filter algorithm continually computes an update of the echo-path model coefficients in order to provide an estimate of the echo that matches the coupled signal as closely as possible. Basically, the algorithm uses the residual error ((e) in Figure 2.19) at the output of the subtracting block for the adaptation process to track variations in the echo-path. The output of the filter results from a sum of signals that are derived from the incoming signal on the receiving path of the echo canceller, which are delayed and multiplied with lower amplitude (a convolution of the incoming signal). Such a filter structure represents a transversal filter with N taps (also called a tapped delay line) with a unit delay between each tap; it is also know as finite impulse response (FIR) filter [1]. The echo canceller has to store the amplitude of every signal for each possible delay between zero and the biggest delay (echo-path capacity) on the cancelled end. The size of N determines the window size of the echo canceller, e.g. for a length of the FIR filter of N = 512 and a sampling rate of 8 kHz the utmost delay of the impulse response equals 64 ms. Because of the filter design, the adaptive algorithm only performs well with linear modifications of the signal between R_{OUT} and S_{IN} (see Figure 2.19). Thus, any non-linear components (e.g., clipped speech segments) worsen the performance of the cancellation process.

The adaptive filter algorithms, which are explained below, are mainly designed for the deployment in a network echo canceller; the main performance parameters are:

- The initial convergence time or speed.
- The tracking performance.
- The divergence behavior in the case of an undetected double talk phase.

Tracking is defined as the ability of the filter to adapt to a change in the echo-path. Normally, a high convergence speed also comes with fast convergence in tracking the local loop of the network. But, at the same time, this behavior also results in a higher sensitivity on distorting signals from the near end. The longer it takes the DTD to notice double-talk periods and to stop adaptation, the more filter divergence is occurring. Network echo cancellers sometimes have the possibility to change to a slower tracking behavior after the filter has reached some convergence state. The reason for this additional mode of operation is that certain hybrid circuits change the behavior very slowly. Of course, this assumption does not hold for acoustic echo cancellation.

The least mean square (LMS) algorithm, which has been previously named stochastic gradient algorithm, and the normalized LMS (NLMS) algorithm are the most widely deployed types of filter algorithms for network echo cancellers. They were first mentioned around 1960 for adaptive switching [52] and at the beginning they were deployed for echo cancellers [48] and adaptive antenna arrays [23]. Later on, the fields of application have also included adaptive signal processing [7] areas like equalization and system identification. On the one side the LMS and NLMS are very easy to implement. On the other side the basic NLMS approach provides only low convergence and tracking speeds. Therefore, several other algorithms have been suggested to improve the performance over the simple stochastic gradient algorithm:

- The proportionate NLMS (PNLMS) algorithm [32].
- Echo canceller based on two echo-path models [45].
- The least-squares (LS) [41] and the recursive least-squares (RLS) algorithm [7].
- The affine projection algorithm (APA) [46].

Beside the NLMS, the PNLMS is the only algorithm that is already implemented in commercial echo cancellers. The step size of the PNLMS used for the adaptation process is proportional to the tap weight. Faster convergence compared to the NLMS is obtained in this way. The obstacles, which the two-path approach has to overcome, are the difficulty in designing a decision algorithm needed for the implementation as well as the required extra memory capacities. The LS and RLS algorithm have not yet been applied widely, because of the high requirement on computation power. Although the RLS algorithm offers a much faster convergence performance, the tracking behavior is not better than that of the LMS algorithm [33]. In fact, both basic types of algorithm, the RLS and LMS, are not well suited to satisfactorily track typical changes in the loudspeaker-room-microphone environment [1]. The APA can be seen as a compromise of both basic types of algorithms, the LMS and RLS, in terms of complexity. Moreover, there are fast recursive least-squares (FRLS) and fast affine projection (FAP) algorithms. Because of the rapid advancement of digital technology, these and other complex algorithms will be used in the near future.

Double Talk Detector

The DTD detects speech periods, in which voice signals appear simultaneously on the sending and on the receiving path of the echo canceller. When the DTD decides on such a double talk situation, it stops the adaptation process of the filter in order to prevent the divergence of the filter. Furthermore, it switches the NLP into transparent mode, i.e. all input signals are transferred unaffected to the output port. More precisely, the DTD reacts on the detection of significant amount of speech from the near end talker with respect to the signals on the receiving path. Thus, double talk mode is even notified, when only cancelled end signals are applied, which is also referred to as single talk at the near end.

The so-called Geigel algorithm [31] belongs to the most widely deployed means of double talk detection. The level based algorithm decides on double talk (see Figure 2.19), whenever

$$|y(n)| > G_{DTD} \max_{n-N_{DTD} \leq m < n} |x(m)|, \tag{2.10}$$

where G_{DTD} is a constant parameter that depends on the expected value of ERL of the echo-path (e.g., $G_{DTD} = ½$ for ERL = 6 dB) and N_{DTD} denotes the number of filter coefficients (also called filter length). The second parameter of the Geigel DTD is the holding time (e.g., T_{hold} = 30 ms). When a double talk period is over, the DTD still holds the output signal for that period of time.

When deployed in a network echo canceller, the Geigel algorithm works reliable, because the ERL value keeps quite constant during a connection. In the case of an acoustic echo canceller the echo-path and, as a consequence, the ERL value may vary considerably; the ERL may even get negative and amplify the echo signal in this way. Recently developed DTD algorithms are based on the correlation [53] or coherence [36] between the signal on the receiving path (x) and the input signal on the sending path (y) or between x and the error signal of the filter (e). The equivalence of both methodologies is shown in [25], as well as a normalized cross-correlation technique is discussed in the same reference.

Non-Linear Processor

The NLP (also named residual echo limiter [21]) blocks low level signals and passes high level signals. The first ones are supposed to be residual echoes from the egress of the adaptive filter and they should be restricted below levels as indicated by the corresponding ITU-T Recommendation [80]. Imperfect cancellation of the circuit echoes occurs due to:

- Non-linear components in the echo-path. As already mentioned, the filter algorithm only handles linear signal reflections adequately.
- Frequency shift of the adaptive filter.
- Finite accuracy of the implemented filter algorithm.

Ideally, the speech from the near end should not be distorted by the NLP. Therefore, the NLP is switched into inactive mode under double talk or near end single talk conditions. The residual echo on the sending path is assumed to be masked out by the louder speech from the near end. Due to unavoidable erroneous detections of the DTD (i.e., the DTD decides on single talk in double talk situations and vice versa), there occur short periods of active NLP in double talk phases impairing the near end speech or the NLP does not suppress residual echoes in transparent mode.

The state of the art in NLP implementations are the widely deployed center clippers (Figure 2.22): Signals that are smaller than the suppression threshold are reduced to zero,

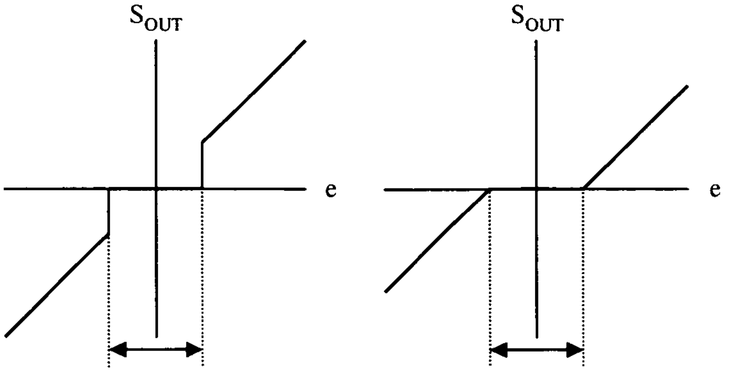

Figure 2.22 – Two examples of NLP characteristics implemented as idealized center clipper transfer functions with adaptive clipping levels.

while signals that are larger than the so-called clipping level are passing without a change. A compromise has to be found between a high threshold preventing high residual echo levels from passing the NLP and low levels to minimize the distortion of the cancelled end speech. Either the clipping level is based on a short-time estimate of the magnitude of the signal on the receiving path or a constant value is taken. In the first variant, the threshold for acoustic echo cancellation, for example, is typically set to a level of about 20 dB below this estimate [1]. The second alternative assumes a certain value for the residual echo level, which is used as a guideline for the suppression threshold.

Under specific network conditions with low delay and/or high ERL values and a sufficient filter attenuation (expressed by ERLE), the NLP can be disabled. As NLPs always introduce some distortions, disabling of the NLP improves the overall voice quality.

Besides, acoustic echo cancellers use more advanced versions of NLP implementations: Soft suppressors have a smoothed transfer function instead of the sudden drop to zero as in the center clipper case. Post processing filters are coming with a frequency selective attenuation [19].

2.3.6 Test Signals

According to ITU-T Recommendation P.501 [100] the following test signals are, in principle, applicable for testing echo cancellers:

- Composite source signals.
- Amplitude modulation (AM) and frequency modulation (FM) based signals.
- White and pink noise.
- Probe tone.
- Simulated Speech Generator: This signal approximates the amplitude distribution of speech through modulating a main signal having a Gaussian distribution by a specially-tailored signal. The resultant is shaped to approximate the long-term frequency spectrum of speech.
- Artificial voice [92].
- Artificial conversational speech [96].

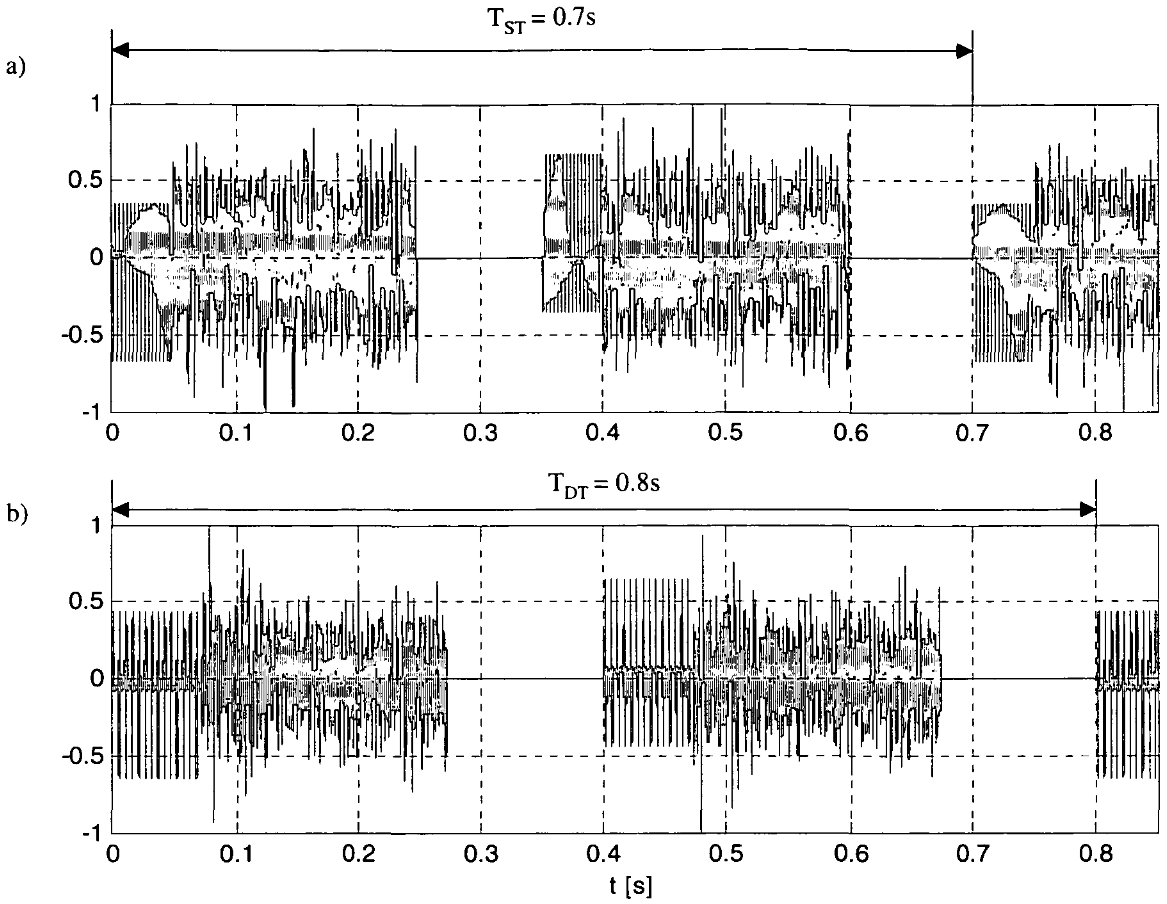

Figure 2.23 – Bandlimited Composite Source Signal in the time domain for a) single talk and b) double talk. Both are built up of a voiced sound, a noise part with speech-like power density spectrum, and a pause.

- Speech-model process controlled by discrete Markov Chains.

Composite Source Signal

The composite source signal (CSS) belongs to the most widely deployed test signals for echo cancellers. As the characteristics of real speech signals depend on various parameters such as language, gender of talker, emotional state, etc., a broad variety of samples would have to be tested in order to obtain reliable results (see also Section 2.1.6). Common attributes of those speech parameters have been extracted and applied on the CSS, which consists of a voiced sound, an unvoiced sequence, and a pause (Figure 2.23). Because of the variety of real speech parameters a common sense for the choice of standardized speech signals would hardly be found. All in all, the CSS gives a direct guideline to the performance of echo cancellers for speech input. The artificial test signal, consequently, offers a convenient and repeatable means of measurement with reasonable effort.

The CSS for single talk and the CSS for double talk are shown in Figure 2.23 a) and Figure 2.23 b), respectively; both differ in the overall period and in the length of the following three sequences:

- The *voiced* signal part is generated from the "artificial voice" signal according to ITU-T Recommendation P.50 [92] and it imitates the voiced parts of real speech signals. It should activate possible speech detectors in voice-controlled systems, which should respond quickly especially to voiced parts of the speech.
- The pseudo *noise* signal is added after the voiced sound in order to provide an adequate measurement signal for the determination of the magnitude transfer function. The time representation is derived from a complex spectrum generated in the frequency domain, which has a constant magnitude with frequency, while the phase is changing [100].
- The *pause* has been included because of two reasons. First, an opening pause puts the communication system with time-varying transfer behavior into a defined initial state. Second, subsequent pauses provide enough time for suitable amplitude modulation to the CSS.

ITU-T Recommendations G.168 [80] and P.501 [100] specify a subset of the composite source signal for testing, the bandlimited CSS for single talk and the bandlimited CSS for double talk. The noise parts of both signals have a speech like power density spectrum with a decrease of 5 dB per octave towards higher frequencies. In order to realize all possible combinations of single and double talk conditions the timely composition and the overall duration of both sequences differs (see Figure 2.23). In the case of double talk, periods of voiced signals applied in both directions as well as situations with voiced and unvoiced signals on both channels are observed, when testing for a sufficiently long period of time. Hence, the correlation between the CSS for single and double talk is kept low. Additionally, the voiced signals are generated with different pitch frequencies and a random noise sequence instead of the pseudo noise signal is applied.

The bandlimited CSS should be applied for all testing with bandlimited systems, which are working in a non-linear way, are time-variant, and need the long-term power density spectrum of typical speech. All these requirements are fulfilled when testing the performance of an echo canceller. In the case of single talk tests, the CSS for single talk should be injected at the receiving port of the echo canceller (R_{IN}). For bidirectional tests on network echo cancellers the double talk sequence is fed in the sending direction at port S_{IN}, while the single talk signal is again applied on port R_{IN}.

AM/FM Modulated Signals

Another methodology for testing echo cancellers are provided by AM/FM signals. The signal generation block for these voice-like frequency composed signals and the resulting speech properties are given in [100]. AM/FM modulated signals are typically utilized for the determination of echo loss, echo loss variation or level variation under double talk conditions. The exact analysis of levels and loss under double talk is based on the separation of the signals after passing the system under test.

2.3.7 Subjective Measurement

A more general introduction on subjective testing has already been given in Section 2.1.6; this part of the thesis concentrates on an applied version of auditive tests designed to assess the conversational voice quality of network echo cancellers. The corresponding standards are ITU-T Recommendation P.831 [107] and G.168 [80], respectively. The collected data and analyzed results of subjective tests are used to extract parameters that determine the transmission quality of speech echo cancellers. Based on these parameters objective test procedures and requirements for laboratory tests are designed.

Auditive techniques are categorized into conversational tests, talking and listening tests, and listening only tests. The assessable parameters for each of the test procedures are given in the corresponding ITU-T Recommendations.

Conversational Tests

Conversational tests are carried out by two subjects communicating over a connection, which includes the echo canceller under test. As the combined subjective effects of all the parameters influencing the conversational quality are evaluated, it is the only way of assessing echo cancellers under realistic conditions. Conversational tests work well for identifying those parameters, which determine the overall quality. But they are not sensitive enough to assess specific characteristics and to make comparisons between different implementations (i.e., the exact test conditions are not repeatable). The tests may, for example, include double talk situations, but the exact starting point in time, the duration and even the number of these episodes are hard to control. Moreover conversational procedures are time consuming and, therefore, expensive. The objective of conversational testing depends on whether untrained or experienced subjects are evaluating different connections. The latter group of persons provides more diagnostic statements.

Talking and Listening Tests

Talking and listening tests are mainly designed to assess talking-related parameters of the echo canceller under (far end) single talk conditions. They require only one subject located at the non-cancelled end, while the other party is emulated by, for example, different echo-path realizations, electrical injection of background noise, or different terminal equipment. Even double talk sequences can be evaluated by using a head and torso simulator (HATS) [95] for feeding in cancelled end speech via an artificial mouth [94]. Moreover, different network conditions and parameter variations of the echo canceller can be tested.

Talking and listening tests are—compared to conversational tests— rather easy to run. Subjects judge either initial convergence or steady state conditions of an echo canceller. Furthermore, they have to perform a special task (e.g., describe the position of given numbers in a figure; also known as Kandinsky test) in order to avoid simple reading. Typically, one of the following parameters is evaluated:

- Disturbances caused by talker echoes.
- Impairments introduced by audible switching.
- Quality of background noise transmission.

Talking and listening tests are better suited for evaluating specific parameters than conversational tests, because the subjects can concentrate better on these parameters, without leading and following a conversation. The test methodology has the drawback that it is more artificial than the conversational tests.

Third-Party Listening Tests

The recording procedure of a third-party listening test requires two artificial head measurement systems such as HATS [107], whereby the echo canceller under test is placed at the far end of the connection (Figure 2.24). Having generated the stereo voice samples, the subjects judge the quality of the recordings via correctly equalized headphones. The subjects perceive the binaural samples as if he or she would participate in the conversation at the non-cancelled end of the echo canceller (i.e., they listen to the conversation as third parties).

Third-party listening tests represent the most sensitive methodology for subjective testing of echo cancellers. They are a specialized version of the listening only test (also see Section 2.1.6) designed for the detailed investigation of various types of speech signal degradations like residual echoes, initial convergence or double talk performance. Based on these parameters, different echo canceller implementations can be directly compared. Therefore, a third-party listening test has been conducted for this work on a new approach of echo cancellation design in order to assess different settings of the echo canceller.

The setup in Figure 2.24 shows an echo canceller on both sides of the connection. Each HATS is equipped with an artificial mouth, two type 3.4 artificial ears [94] and mounted handsets. The artificial mouths are both fed with pre-recorded source material. If double talk sequences are evaluated, a male and a female voice are induced with the intention to provide clearly distinguishable talkers for the listeners. The setup reproduces important characteristics perceived by a real-world talker like the acoustical leakage between handset and the human ear, sidetone and self-masking. More detailed, the recordings of the binaural samples made at the far end side (according to the echo canceller under test) consist of:

- The far end voice coupled from the artificial mouth to the open ear microphone (not covered by the handset).
- The distant voice coupled from the artificial mouth to the ear microphone covered by the handset via the handset sidetone and the acoustical leakage.
- Echo signals from the voice originated at the non-cancelled end transmitted via the receiving direction of the terminal.
- The double talk signal (induced at the near end) transmitted via the receiving direction of the terminal.

The last three transmission characteristics in this list are pressure force dependent and thus standardized [94]. The assessment of the samples may be carried out by untrained or trained subjects depending on the objective of the test.

There are various advantages of a third-party listening test compared to the two test procedures explained above:

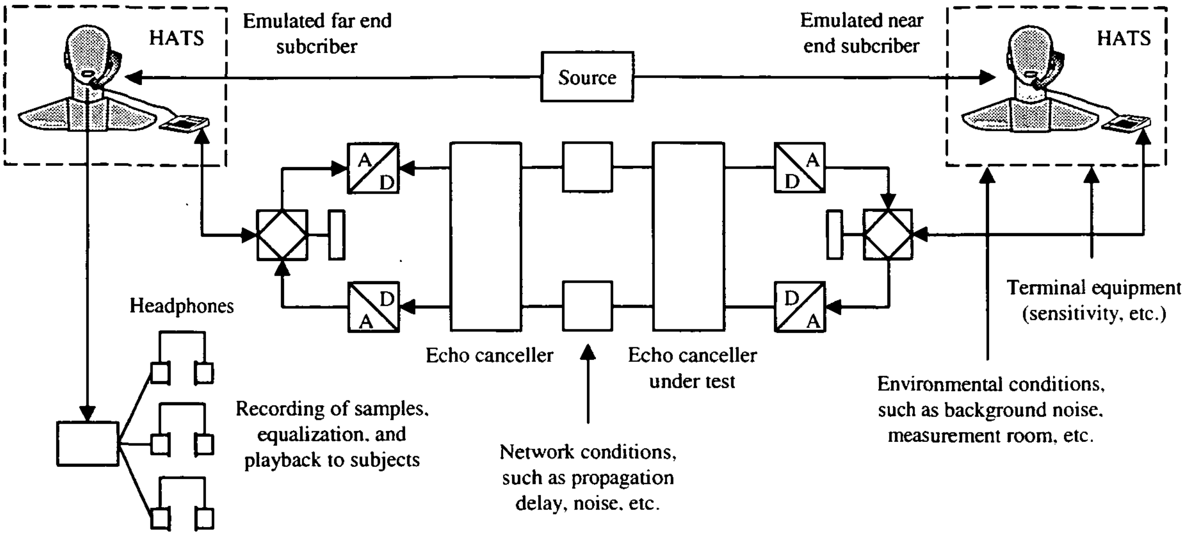
HATS
Emulated far end subcriber
Source
Emulated near end subcriber
HATS
A
D
D
A
A
D
D
A
Headphones
Echo canceller
Echo canceller under test
Terminal equipment (sensitivity, etc.)
Environmental conditions, such as background noise, measurement room, etc.
Recording of samples, equalization, and playback to subjects
Network conditions, such as propagation delay, noise, etc.

Figure 2.24 – Setup for recording of binaural speech material for a third-party listening test using two HATS. The listeners use equalized headphones for the assessment.

- The test conditions in the setup are easy to control. That's why, different echo cancellers can be tested under identical settings and the number of measurement conditions is varied with no difficulties.
- The test procedure is less time consuming when including further echo cancellers, implementations or environmental conditions.
- The number of subjects is made larger in an easy way, and the recordings have to be done only once.
- The overall test procedure may be divided into smaller parts with specific attributes to test.
- The recordings with two artificial head measurement systems can be done for single and double talk situations.
- The test is capable to judge specific attributes because subjects are able to concentrate on these more thoroughly.
- The recordings provide a very high level of realism for the third-parties evaluating the speech material. The perception of the sounds is determined by various parameters like sensitivity, linear and non-linear distortions of terminal equipment, coupling between handset and ear (leakage), handset sidetone, masking effects and others. The listening samples are recorded at the acoustic interface, thus offering all the parameters revealed above are considered.
- The test allows direct A/B comparison of echo cancellers [104].
- Because of the high sensitivity of the test methodology even small variations between different echo cancellers can be judged.

A fundamental drawback of the third-party technique is that the subjects involved are not allowed to talk. They listen to and evaluate recordings of unknown speakers and do not perceive their own voice. On the other hand, this type of listening test provides a very efficient procedure to assess differences in various implementations of echo cancellers.

Selection of Subjects
Subjects have obtained different experiences concerning the subjective testing and the knowledge about echo cancellers. Basically there are untrained and experienced subjects.

Untrained subjects are familiar with the daily use of a telephone. They have neither knowledge about the operation and maintenance of echo control devices nor have they been participating in subjective tests so far.

Experienced subjects have already come into contact with subjective testing, but they do not regularly carry out assessments. Moreover, they are able to give a detailed description of an auditory event and of the corresponding subjective impression. They are capable of distinguishing diverse events according to the specific distortion. Experienced subjects neither know how particular implementations influence the perceived quality nor do they have competences on the field of technical implementations of echo cancellers.

Analysis of results
In order to derive reasonable results from the single voting of each subject, it is necessary to apply standard statistical procedures, as noted in ITU-T Recommendation P.800 [104] and in the Handbook on Telephonometry [122].

2.3.8 Objective Measurement

Objective test procedures for echo cancellers measure attributes of the transmitted signals in terms of echo levels, echo delay, background noise levels, spectral compositions, time characteristics, etc. under single and double talk conditions. The human perception is not taken into account; such a method would have to consider all talking and listening related attributes of the conversation. Within ITU-T Study Group 12 an enhanced version of the objective speech clarity measurement method PESQ for the evaluation of echoes and sidetones has been contributed [121]. A competition was scheduled for 2003 on talker quality, but the corresponding topic has been removed from the work program of the study group [173]. As the contest has been canceled, no measurement method has been standardized so far.

Proprietary methods for the objective evaluation of the voice quality have been implemented in some measurement devices. They just take an objective speech clarity measurement method (e.g., PSQM), which has been exclusively designed for pure listening and not for talking situations, and apply the original talker's voice as reference and an artificial signal as disturbed sample. The latter is artificially created by the measurement device by superposing the reflected echo signal with the undisturbed reference. Furthermore, no time-alignment between reference and transmitted signal is applied. The Agilent VQT [168], for example, applies an enhanced version of the PSQM method on both signals and introduces a new measure, the so-called perceived annoyance caused by echoes (PACE) for the resulting voice quality. Such a procedure must not be utilized for the measurement of talker echoes, as—among others—the masking of the received signal through sidetone is not considered and perceived voice quality algorithms have not been designed for such applications.

The main sources of test and measurement methodologies for network echo cancellers are ITU-T Recommendations G.168 [80] and P.502 [101] as well as a specification from Deutsche Telekom [20]. The latter describes procedures, which are similar to ITU-T Recommendation G.168, but the guideline provided for the critical double talk test is better suited. Basically, echo cancellers should fulfill the following fundamental requirements [80]:

- Fast convergence.
- Low returned echo level in far end single talk situations.
- Low divergence rate during double talk and during near end single talk.
- Reliable double talk detection and cancelled end speech detection.
- Undisturbed transmission of fax and low speed (< 9.6 kbit/s) voice-band data signals.

Corresponding measurement procedures inject composite source signals, noise, tones, fax signals, and voice-band data signals as test signals. The following procedures represent the main tests for echo cancellers: Double talk test, leak rate test, infinite echo return loss convergence test[24], stability test, comfort noise test, fax test, and performance with low-speed data modems.

2.3.9 Voice Quality Impairments

Various subjective and objective tests of echo cancellers have pointed out that the voice quality is mainly determined by the performance under double talk conditions. As already mentioned before, the design of the NLP implementation together with the robustness of the DTD is extremely critical. The occurrence of cancelled end speech or a double talk situation must be detected by the DTD in order to control the NLP properly. Imperfect detection of double talk combined with a high suppression threshold level of the NLP has the consequence of an impaired near end speech. The echo canceller, then, shows some of the characteristics of an echo suppressor. The lower the suppression level, the easier are double talk signals passing the echo canceller—even if the DTD falsely detects on far end single talk—because the non-linear distortions in the cancelled end speech are kept quite low. But if the NLP level is too low, then peaks of the residual echo may traverse the NLP and disturb the talker at the far end. Due to the imperfect functioning of the DTD the following aspects concerning the voice quality are stated:

- The most annoying effects occur, when the NLP is inserted during continuous speech, due to an erroneous decision on single talk in a double talk situation. The resulting speech gaps have a more degrading influence on the perceived quality than front-end clipping at the beginning of a double talk sequence [63]. The durations of these temporal suppressions of the double talk signal should be kept lower than 64 ms. In this range the amount of clipped speech should not be greater than 0.1 percent of time [80].
- Furthermore, the transmission of the background noise present at the near end sub-

---

[24] Pure four-wire connections, such as ISDNs, are regarded as echo-paths with infinite ERL.

scriber's location may also be suppressed, which degrades the perceived speech quality significantly [40] [63] [80].
- The DTD always needs some time to notice double talk periods, resulting in a delayed change of the NLP mode into double talk operation (i.e., transparent NLP behavior). In the same time period the adaptive filter diverges due to near end speech. The improper NLP mode affects the near end speech and the diverging filter causes higher residual echo levels [62] [76], which may to some extent be masked by the voice originated by the user on the cancelled end.

Further results of subjective conversational tests for commercial echo cancellers are published in [118]. Results of echo cancellers implemented in VoIP gateways for carrier solutions and provided by different manufacturers are published in the anonymous test report of the 2nd ETSI speech quality test event for VoIP equipment held in April 2002 [155]. The tests focused on double talking and they were conducted with level-varying composite source signals (designed for single talk and double talk, respectively) in sending and receiving direction [101]:

- Even in a complete four-wire connection (which is, in fact, free of echoes) most implementations turned out to significantly impair the perceived voice quality of the distant subscriber. The disturbing effects comprise level variations and signal clipping.
- Echo-path realization with ERL = 40 dB lead to similar results compared to the open echo-path test condition.
- In addition to level variations and clipping the echo canceller also returned significant echo levels for the minimum ERL value of 6 dB.

Additional measurements during double talk were carried out using the AM/FM modulated test signal described in Section 2.3.6.

2.4 Conclusion

As summarized in the preceding section, the NLP of network echo cancellers is responsible for disturbances due to erroneous decisions of the DTD. On the one hand, when notifying single talk in a double talk situation, the center clipping implementation of the NLP with adaptive thresholds (see Section 2.3.5) cuts away parts of the voice signals (i.e., the NLP introduces non-linear distortions such as clipping and speech gaps). Such NLP realizations are often found in practice and typically completely suppress the residual echo under far end single talk conditions. On the other hand, in the case of single talk originated at the far end and—at the same time—switching the NLP into double talk mode (i.e., the NLP is made transparent from the signal point of view although it should perform as center clipper) results in unchanged echo signal reflections for the talker. The resulting annoyance of talker echoes always depends on two parameters, the echo attenuation and the delay experienced from the talker's mouth to the talker's ear. High levels of echo delay, for example, do not disturb the talker, if they are below 10 ms.

Since IP based telephone systems are always coming with a relevant amount of delay (see Section 2.2) and, moreover, the echo cancellation implementation in high-end IP solutions for carriers may impair the near end signal especially in double talk periods (see Section 2.3.9) the future success of IP telephony will strongly depend on the performance of the echo control device.

Chapter 3

Delay-Controlled Echo Cancellation

> "A physician can sometimes parry the scythe of death, but has no power over the sand in the hourglass."[25]

3.1 Variable NLP Attenuation

To overcome the impairments of the conversational voice quality, which are caused by standard network echo cancellers, the new concept on echo cancellation replaces the center clipping NLP with a delay-controlled attenuation. Based on the present echo delay (i.e., the round-trip time experienced by the talker echo on the way from the talker's mouth to the talker's ear) and the current attenuations along the same path, the NLP provides a sufficient, delay-dependent but limited additional attenuation in order to reduce talker echo levels below values as indicated by the "acceptable" curve in Figure 2.16 [26]. The perception oriented, delay-controlled echo canceller relies on the assumption that echoes do not necessarily need to be completely suppressed for avoiding disturbances. In this way speech quality degradations inserted by the NLP are minimized even when the DTD wrongly turns

---

[25] Hester Lynch Piozzi Thrale (1741–1821), British writer. Letter, November, 1781, to author Fanny Burney.

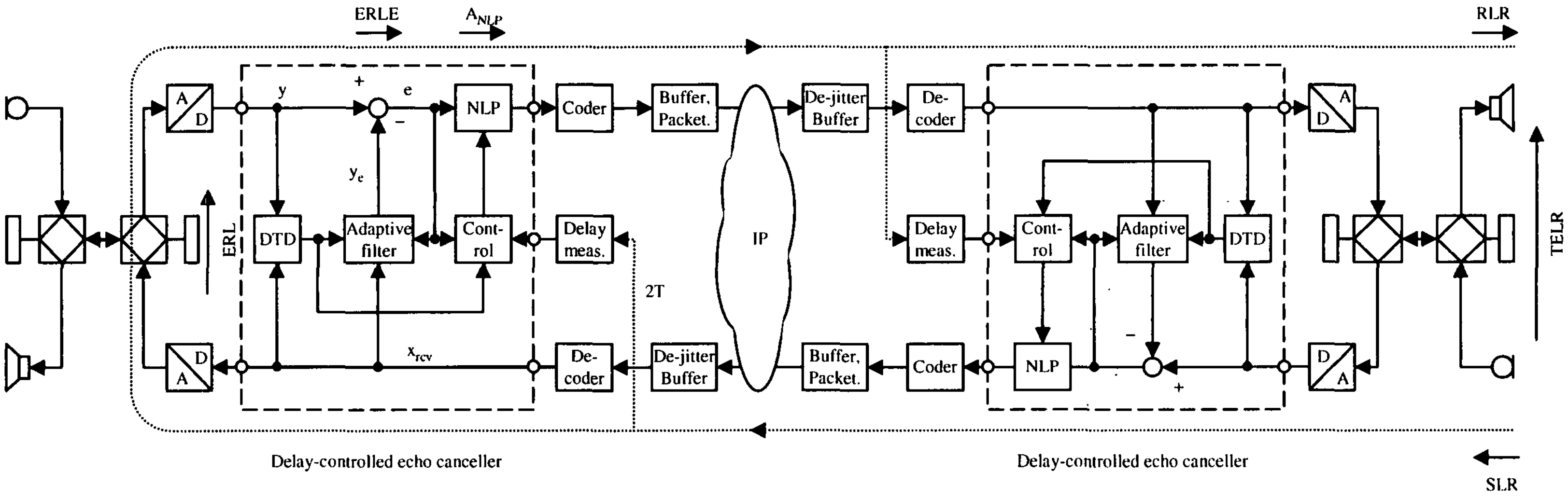

Figure 3.1 – The new concept on echo cancellation deployed in a VoIP network, which connects two PSTNs, requires two measures for controlling: The round-trip delay 2T and the corresponding attenuations along the overall echo path from the talker's mouth to the talker's ear.

on the NLP in a double talk situation (where it should transfer the input signals unchanged to the output). The new approach improves the conversational quality only in such undetected talking periods, while in the case of single talk at the far end the NLP allows a certain amount of echo, which should be almost unperceivable by the corresponding talker.

The room of improvement for the voice quality of the near end speech increases with low echo delays, because the required TELR values and—as a direct consequence—the required attenuation introduced by the NLP are reduced. With decreasing NLP attenuation, moreover, the new idea differs even more from a center clipper implementation, which should provide residual echo levels below -65 dB_{m0} independent from the current delay [80]. On the contrary, with increasing echo delays the behavior of the perception oriented, delay-controlled concept approaches the conventional one.

Furthermore, the new approach is applicable independent of the implemented filter algorithm or post processing component (e.g., linear attenuation, level switching device, Wiener Filter [7]). Beside the application on network echo cancellers, the new approach is also implementable in acoustic echo cancellers. Due to typically lower filter attenuations provided by acoustic echo cancellers, a more aggressive NLP implementation is needed in order to sufficiently suppress residual echoes. This may lower the effectiveness of the concept compared to electric echo cancellers, but improvements are still expected [26].

Compared to the basic idea on network echo cancellation illustrated in Figure 2.19, both delay-controlled devices in Figure 3.1 are additionally equipped with a delay measurement unit, which continually provides the newly designed control entity with mouth-to-ear delay values (i.e., with every received IP packet) or at certain points during a connection (e.g., only once before call setup or when the echo delay changes to a certain amount). Corresponding measurement methodologies are discussed thoroughly in Section 3.2.

The attenuation introduced by the NLP (A_{NLP}) and calculated by the new control unit results from the current hybrid and from the present adaptive filter attenuation—the ERL and the ERLE, respectively (see Figure 3.1):

$$A_{NLP} = \mathrm{TELR}(T) - (\mathrm{ERL} + \mathrm{ERLE} + \mathrm{SLR} + \mathrm{RLR}), \tag{3.1}$$

whereby A_{NLP} must not be negative, as this would be the same as an undesired amplification of the echo signal. Furthermore, assumptions for the SLR and RLR values of the deployed phones have to be made in Equ. (3.1). These values characterize the sensitivity of the terminal (see Section 2.3.3) used at the far end side. As the delay-controlled echo canceller improves the achievable voice quality especially in IP environments with low delay values, possible applications are, for example, found in corporate or in-house telecommunication networks. In most cases such systems are equipped with one common telephone device and, thus, the system engineer knows the exact SLR and RLR values. However, standardized values and their tolerances for European and North-American countries are given in Table 3.1.

The worst case scenario for "SLR + RLR" in Table 3.1 would assume a sum of 3 dB and 2 dB according to ETSI and ANSI, respectively. The sum of ERL and ERLE in Equ. (3.1) equals the overall attenuation of the residual echo present at the output of the differencing unit (e) in terms of the signal on the receive path of the echo canceller (x_{rcv}) excep-

	SLR [dB]	RLR [dB]	Tolerance [dB]
ETSI TBR 8 [150]	7	3	± 3,5
ANSI TIA/EIA-810-A [156]	8	2	± 4

Table 3.1 – Recommended SLR and RLR values as well as the corresponding tolerances.

tionally under far end single talk conditions [1]:

$$\mathrm{ERL} + \mathrm{ERLE} = 10\lg\frac{\mathrm{LPF}\left\{x_{\mathrm{rcv}}^{2}\right\}}{\mathrm{LPF}\left\{e^{2}\right\}}. \tag{3.2}$$

A time-discrete realization of the low-pass filter (LPF) has been used within this work for the creation of the voice samples for the third-party listening test (see Section 4.2). As the NLP control unit evaluates Equ. (3.2) only in the absence of near end speech and as it holds the combined attenuation along the echo-path and adaptive filter in double talk periods, the uncertainty of the resulting sum relies, among others, on the DTD implementation. Even small delays of about a few milliseconds in detecting such phases results in considerable divergence of the filter algorithm.

All in all—having measured the current echo delay—the computation requirement for Equ. (3.2) is kept quite low, since the table look-up operations for TELR values in terms of the echo delay are simple, and only basic operations for the sum of ERL and ERLE according to Equ. (3.2) are needed.

Some examples for required NLP attenuations under varying echo delay conditions are given in Table 3.2. The calculation of A_{NLP} assumes the minimum and, thus, worst case hybrid attenuation of ERL = 6 dB [80], SLR + RLR = 10 dB (see Table 3.1), and a constant filter attenuation of ERLE = 20 dB.

Moreover, the new approach is easily extended for the consideration of masking during double talk, which comes with a lower sensitivity of echo perception compared to far end single talk situations: Instead of switching the NLP into transparent mode under double talk conditions (A_{NLP} = 0 dB) and, at the same time, allowing unaltered passing of echo signals, the attenuation provided by the NLP in single talk situations is reduced taking into account the masking of the far end talker echo by the double talk signal. This measure is taken, because the user at the far end side accepts a higher level of echo, when signals are

T [ms]	15	30	50	90	160	300
TELR [dB]	28,2	34,9	40,1	45,8	50,6	54,6
A_{NLP} [dB], Note 1	0	0	4,1	9,8	14,6	18,6

Table 3.2 – Examples for different echo delays, the resulting TELR values, and, finally, the required values for A_{NLP} are based on the following assumptions: ERL = 6 dB, ERLE = 20dB, and SLR + RLR = 10 dB. Note 1: The computed A_{NLP} values for 15 ms and 30 ms are made zero, because A_{NLP} must not be negative. In those cases the NLP does not influence the signal to be transmitted.

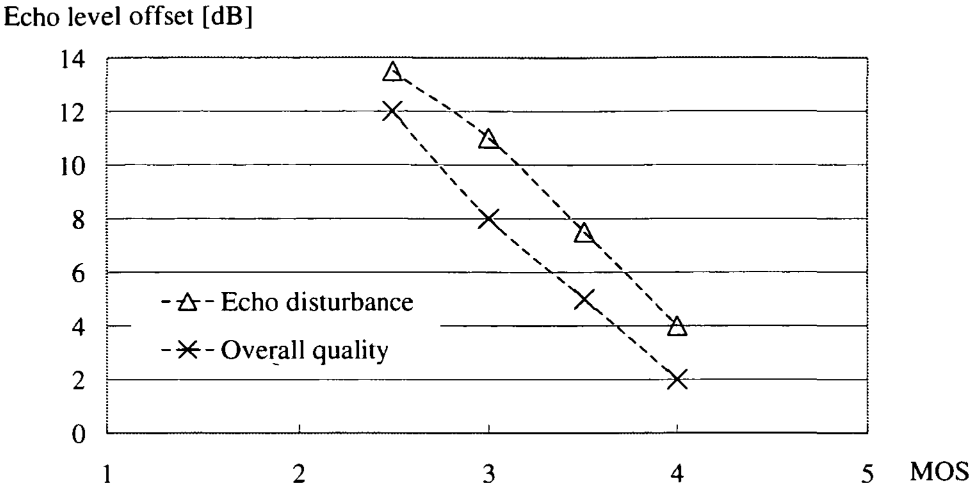

Figure 3.2 – Higher echo levels are allowed during double talk to achieve the same MOS values compared to single talk conditions. The results were obtained in subjective tests for hands-free telephone situations with a one-way transmission time of 100 ms.

applied simultaneously in both directions of the echo canceller. The difference between the TELR values in single talk periods and the corresponding values under double talk for the same level of voice quality in a hands-free situation are given in Figure 3.2 [76] [120]. The parameters overall voice quality and echo disturbance were assessed in third-party listening tests. Although the graphs in Figure 3.2 were obtained in a hands-free telephone situation with a one-way delay of 100 ms, it is assumed, that there are similar relations for other configurations with a gain of at least a few decibels in terms of TELR values. To include

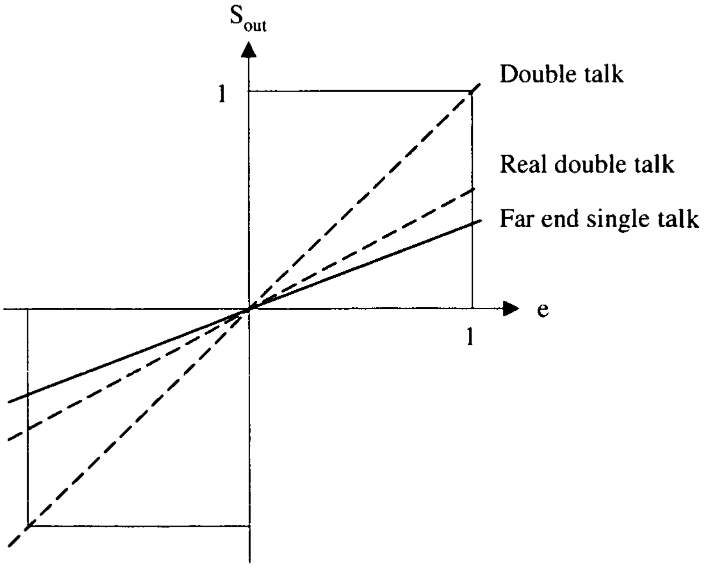

Figure 3.3 – Linear characteristic curve for NLP implementation with slightly reduced (2 dB) attenuation under real double talk conditions—compared to single talk situations.

the masking during double talk into the control mechanism of the new approach, the DTD has to notify another special talking situation, the *real double talk*. As already mentioned in Section 2.1.5, the DTD also decides on double talk, when there is significant amount of speech exclusively present at the near end and, i.e., that there is no signal occurring at the incoming path of the echo canceller. The real double talk operation mode is notified by the DTD, when remarkable voice signals are present in both directions of the echo canceller at the same time.

In its simplest version the NLP is realized by a linear attenuation as shown in Figure 3.3. The dashed line of the newly introduced case of real double talk shows a higher slope (e.g., plus 2 dB) than the far end single talk mode, as higher echo levels are masked by the double talk signal.

The determination of the mouth-to-ear talker echo delay represents the most challenging task within the new idea on echo cancellation for IP based telephone networks.

3.2 Delay Measurement

The reliability of the NLP control mechanism draws from accurate determination of the echo delay. Basic definitions as well as possible sources of errors and uncertainties have been discussed in Section 2.2.3.

Basically, the new approach comes with a theoretical limitation concerning the delay measurement: The NLP attenuates residual echo signals according to the expected round-trip delay along the echo path *before* they experience the delay in the sending direction. Though the control mechanism has knowledge about the network delay on the receiving path, it is, of course, unaware of the expected delay in the sending direction. Therefore, the corresponding one-way delay values are estimated from preceding packets on the transmission path. Thus, it is highly recommended to monitor both one-way delays continuously in order to keep the resulting errors low.

The required precision of the delay determination, expressed by the sensitivity of the one-way delay T in terms of the overall attenuation along the echo path TELR, is derived from the inverse slope of the "acceptable" curve illustrated in Figure 2.16. For this purpose both talker echo tolerance curves are redrawn in Figure 3.4 with a linear time scale by using Equ. (2.7). As both graphs in Figure 3.4 are displaced by a fixed amount of 6 dB, the numerical solutions (dT/dTELR) for both trajectories are identical. Moreover, the derivatives are constantly ascending, i.e., when measuring in the high echo delay range the requirements on the delay determination are decreasing and vice versa. When assuming, for example, $T = 300$ ms or $T = 30$ ms, and a maximum imprecision for the level measurement of 1 dB, the required delay values have to be within 57 ms and 3 ms, respectively.

Basically, the measurement of the echo delay from the talker's mouth to the talker's ear is based on one of the two following concepts:

- The evaluation of the life traffic or an explicit measurement signal provides the control mechanism with round-trip delay values (see Section 3.2.1).

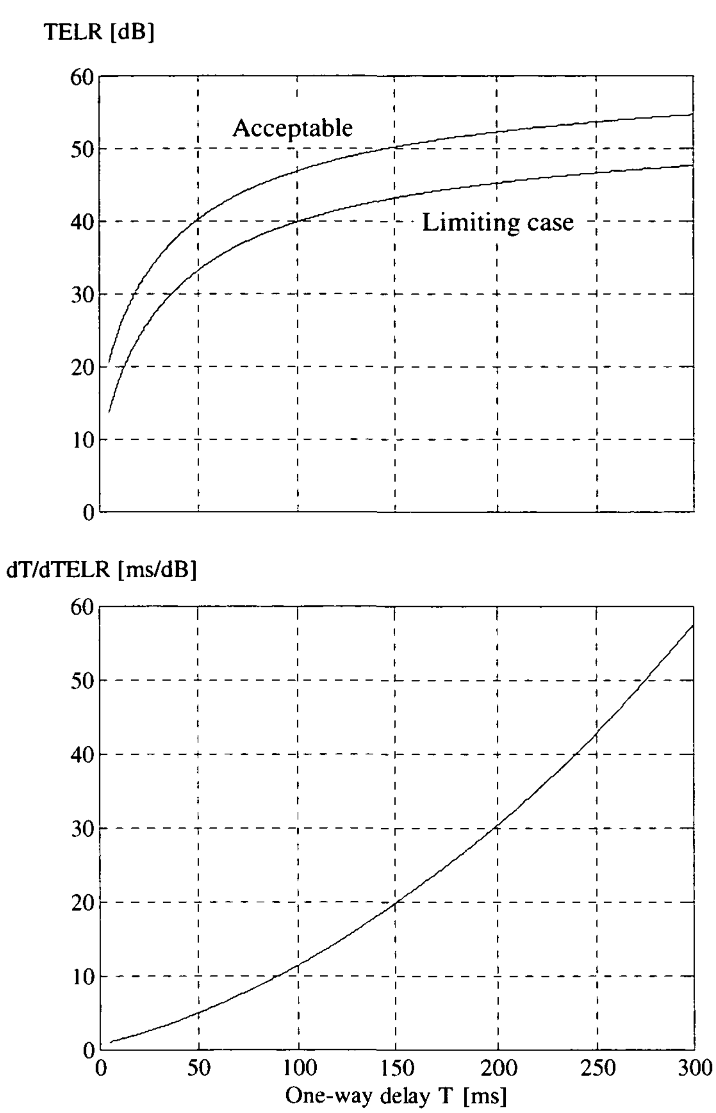

Figure 3.4 – Talker echo tolerance curves based on a linear one-way transmission time scale and the corresponding inverse slopes, which are identical for both curves.

- The timestamps of the IP packets enable the calculation of the one-way delay and with some modifications also the direct evaluation of the round-trip delay experienced through the network.

3.2.1 Signal-Based Measurement

There are two ways how to assess a measurement signal: Either the reflected life traffic allows an implicit determination, or an explicit impulse exclusively inserted for test purposes makes the quantification of the required echo delay possible.

Implicit Methodology

The new approach on echo cancellation always comes with residual echoes reflected to the subscriber placed at the non-cancelled end. The talker echo delay is assessable especially under far end single talk conditions by building the correlation function of the echo signal with the originally transmitted talker signal. The echo canceller has to record the samples leaving the egress port into the sending direction over a sufficiently long period of time (e.g., up to a window size of 300 ms) for this purpose.

Two adversely situations impede significant peaks in the correlation result: Firstly, in double talk situations the echo signal may not be sufficiently distinguishable to the double talk signal. Secondly, in the case of high attenuations along the echo path, the received echo level at the delay measurement unit may be too low for an accurate estimation. But, in the latter situation, the user may also not be able to recognize echo. All in all, under both conditions the control unit adds some uncertainty to the measured value in order to correct the occurring error.

The continuous monitoring and the low computational effort are advantageous for the implicit evaluation of the life traffic.

Explicit Methodology

The intrinsic injection of an explicit test signal enables the evaluation of round-trip delay values experienced along the talker echo path. The assessed values are taken as estimates for the echo delay. As this approach has a strong influence on the network behavior, it should be conducted only once during a call (e.g., in the call setup phase) by both echo cancellers. Therefore it is regarded as more theoretical idea within this work. The delay measurement unit injects the test signal, which is arranged in standard real-time IP packets, into the IP network. Figure 3.5 illustrates the measurement path—represented by the inner dotted line—assessed by the echo canceller placed on the left-hand side.

The result of the active measurement includes all IP specific signal processing components—the far end delay-controlled echo canceller, and the corresponding echo-path. On

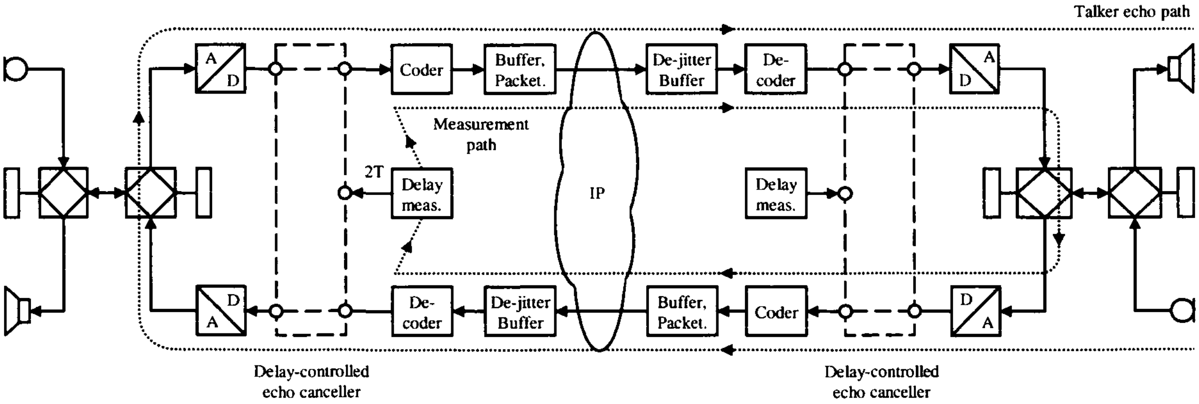

Figure 3.5 – An explicit measurement signal emitted by the delay measurement component of the left-hand echo canceller traverses the speech path and provides a one-time estimate of the round-trip echo delay. The same applies for the echo control unit at the right-hand side.

the contrary, the required near end hybrid and the customer loop attached to it are not included in the evaluated path. Usually, the telephone of the near end user and, thus, the local loop do not introduce considerable amount of delay. But there are also some exceptional situations: When there is another telephone network attached to the gateway (which incorporates the echo control device) or when special equipment such as, for example, a digital enhanced cordless telecommunications (DECT) system is used, the overall echo delay increases significantly. As the new idea on echo cancellation is designed for closed communication system, the deployment of such systems is under the control of the systems engineer and, thus, foreseeable.

Therefore, one essential thing has to be pointed out in this context: The echo control unit situated at the left-hand side in Figure 3.5 has to compensate echoes originated by the left end hybrid—the corresponding talker echo path is represented by the outer dotted line. There are some differences compared to the measured path: Firstly, the measurement takes into account the far end hybrid, which is not traversed by voice signals from the distant party. Usually, there is no remarkable difference in terms of delay over different hybrids in the same network. However, separate measurements should be conducted in order to ensure delays of the same size. Secondly, the echo canceller carrying out the explicit delay determination has to consider its own delay introduced by signal processing. Generally, the group delay in the sending path of a network echo canceller should not exceed 1 ms. Moreover, the receiving path must not introduce notable delay [80]. Therefore, the additional delay is quite low and known in advance. Finally, the delay experienced along the local loop of the far end subscriber is not considered, as well. Its influence on the overall delay is negligible.

Basically, there are two types of measurement signals, which are used for the determination of round-trip delay values:

- A test signal composed of a high, single peak with high slopes.
- A disabling tone.

The first impulse must not be removed by the remote echo canceller, even if this device is in a non-convergent state with a high threshold of the center clipping NLP. A verification of this behavior has to be accomplished in the field of application. The far end echo control device is not turned off during the evaluation and performs as if it would perceive standard voice signals. The second signal corresponds to a 2100 Hz tone with or without phase reversals and it is detected by the so-called *tone disabler* (also known as tone detector) of the distant echo canceller. The following characteristics of the disabling tone and the tone disabler are written in the corresponding ITU-T Recommendation [80]. Every echo canceller should have implemented such a function. Disabling should not take place upon detection of other in-band signals (e.g., speech). If phase reversals are occurring in the disabling tone, the whole echo canceller should be disabled, while if they are missing, only the NLP should be turned off. The latter case occurs, when fax signals and low-speed voice-band data are transmitted. The term disabled in this section refers to a condition in which the echo canceller does not modify signals which pass through it in either direction. Under this condition, echo estimates are not subtracted from the signal on the sending path, the nonlinear processor is made transparent, and the delay through the echo canceller still meets the specified conditions. Moreover, the tone disabler should detect and respond to the disabling signal, when it is present in either the sending or the receiving path. The desired round-trip delay experienced from the sending to the receiving path is determined according to [80]:

> " [..] it is possible, for example, to measure the round-trip delay of a circuit with the disabling tone but the trailing edge of the tone burst should be used and sufficient time for all devices to be disabled should be allotted before terminating the disabling tone and starting the timing."

The frequency characteristics of the tone detector are given in Figure 3.6. The tone disabler must detect tones in the frequency range of 2100 Hz ± 21 Hz, while in the band between 1900 Hz and 2350 Hz detection may occur. The corresponding input levels reach from 0 dB_{m0} down to -31 dB_{m0} and -35 dB_{m0}, respectively. Furthermore, periodic phase reversals, if any, have to occur every 450 ± 25 ms. Phase variations in the range of 180° ± 25° should be detected, while those in the range of 0° ± 110° should not be detected. This restriction has the purpose to minimize the probability of false disabling the echo canceller due to speech currents and network-induced phase changes. Further details of the disabling signal are defined in [112] and [113]. Interferences with other signalling systems are avoided by using only the recommended and internationally deployed disabling tone.

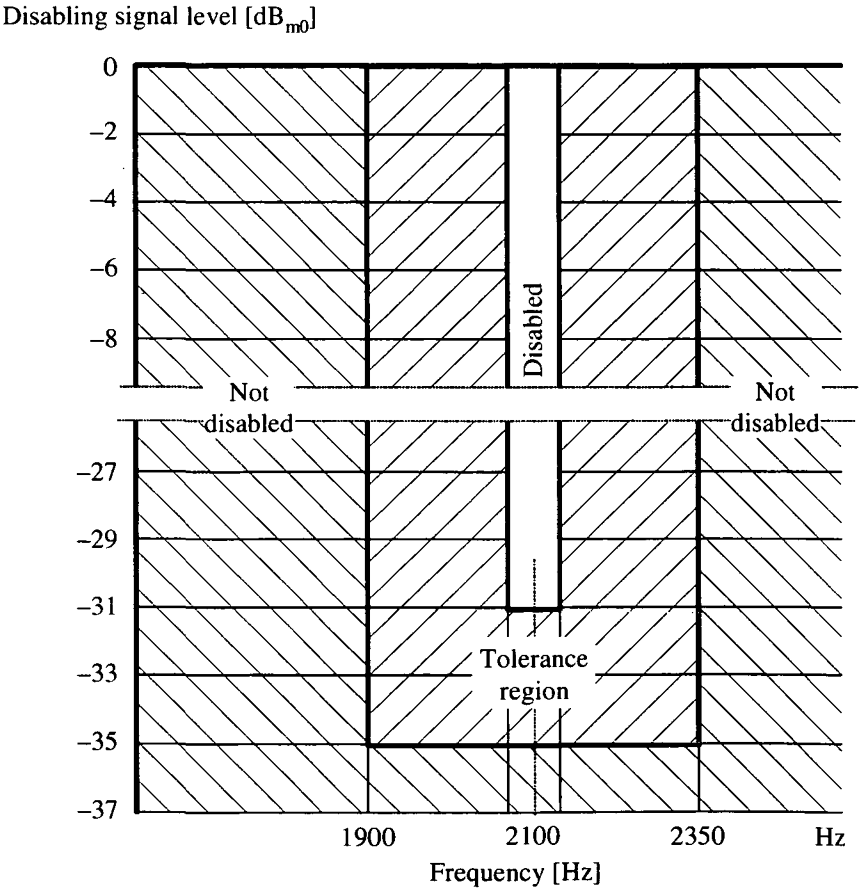

Figure 3.6 – Required band characteristics for the disabling tone.

Finally, the guard-band and holding-band characteristics, the operate time, false operations due to speech currents and data signals, and the release time are described in detail in [80].

Siemens Digital-Echokompensator DEC

Beside the standardized means of disabling the echo canceller and measuring the echo round-trip delay, some vendors also provide further instruments to switch off the echo control device. Representing most features of commercially available network echo cancellers, the Siemens digital echo canceller (DEC) [21] disables some or all functions (i.e., only the NLP or even the adaptive filter) upon the detection of one of the following incidents:

- Detection of bit "a" or bit "c" in the D-channel, which are sent by the switch during call setup in the case of data transmission or cascaded transmission links. The concept of setting one of those bits in the signalling channel is recommended by the Deutsche Telekom in [20].
- Notification via high level data link control (HDLC) D-channel protocol: Echo cancellers look at exchanged protocols between the involved switches. Evaluation of HDLC bit patterns is mainly realized in private networks.
- A 2000 Hz receiver for testing of transmission links via SS7. The far end switch sends back a 2000 Hz signal to the sender. The echo canceller has to be turned off during transmission.
- Receiving of a pre-defined bit pattern over a certain period of time.

As recommended in the corresponding ITU-T Recommendation, the DEC switches off the NLP function, when a 2100 Hz disabling tone without phase reversals is detected. At the same time the adaptive filter still delivers echo estimates. On the contrary, when periodic phase reversals are occurring additionally, the complete echo canceller is disabled. The latter function is implemented for the measurement of long distance connections, where exact results are only achieved with disabled echo control devices.

The fundamental drawback of an explicit measurement is that the additional test traffic disturbs the network performance. Even if it is possible to inject and measure the specially formed signal unnoticed by the user, the round-trip delay values are obtained only at certain points in time during a call and, thus, not continually. The only reliable way of measurement is to carry out the determination before the participants start communicating. Changes of the network behavior are not tracked sufficiently in this way. Other methodologies providing delay values in a continuous way are discussed in the following section.

3.2.2 Timestamps

The frames at the output of the codec, which contain digitized and compressed voice of 10 to 30 ms length, are packed into IP datagrams, and sent under the control of the end-to-end transport protocol RTP (see Sections 2.1.2 and 2.2.1). In order to play out the sequences at the receiver in the right time pattern, each packet is provided with a timestamp added just prior to the sending of the first bit. The receiving timestamp is evaluated immediately after the arrival of the whole datagram. Consecutive frames are time-aligned in this way.

Timestamps are able to provide the NLP control mechanism with network delay values for every received IP packet. According to Section 2.2.1, the network delay comprises the serialization, the queuing, and the propagation delay (i.e., the time span from placing the first bit on the network until the last bit is arriving at the receiver). The coding delay, which includes the packetization delay, as well as the de-jitter buffer delay are not considered in the network delay and, thus, have to be added to the mouth-to-ear delay afterwards.

Round-Trip Time

Before introducing the determination of the one-way delay based on synchronized clocks, a novel concept for direct and continuous assessments of the overall network delay experienced in both directions (i.e., the pure network round-trip delay) is discussed—it is carried out without synchronization.

When the echo canceller receives a voice packet, it takes the original sending timestamp of that packet as well as the receipt timestamp and places both in the next sending packet (which, of course, also contains the new sending timestamp). Therefore, the overall number of received timestamps always amounts to three. Finally, the receiver evaluates the round-trip time by using a fourth timestamp, which is obtained upon the arrival of the packet. The reply delay at the echo canceller is made assessable in this way and it is substracted from the overall transmission time in order to derive the correct round-trip value. All in all, the sending timestamps arriving at the echo canceller are sent back to its origin with this new methodology. The overall procedure corresponds to the NTP time inquiry mechanism, which is discussed later on in this section. Moreover, a modified and enhanced (and not standardized) version of the RTP is required for this new means of continuous round-trip delay measurement.

As the number of transferred packets in both directions of a communication may vary considerably especially in the case of speech pauses and activated VAD, some of the arriving timestamps have to be dropped, because only the most recently received time information is placed in the next sending packet. If packets containing the reflected signals are looped back to its origin, time values for the control mechanism are only needed in far end single talk situations, since echoes are only perceived by the distant talker.

The new concept on continuous measurement of network round-trip delay is based on an enhanced version of the RTP (two additional timestamps are placed in the protocol header) and it is implementable without synchronization.

One-way delay

The standard methodology of monitoring one-way delays continuously requires synchronized receiver and transmitter clocks. The corresponding values are derived from the differences of the receiving and sending timestamp, which are found in the RTP header. Absolute accuracy (i.e., the difference to UTC time) has, of course, no influence on the measurement result. The only parameter of interest is the difference of the current system times on both sides of the connection, which is referred to as synchronization.

To achieve synchronization, on the one hand, both parties gather time information from selected time servers in the network by utilizing the NTP. On the other hand, the use of satellite receiver cards guarantees time information with very high accuracy. The basic concept of both methodologies for exchanging time information have been discussed in Section 2.2.3. The following describes the adaptations of both methodologies to meet the requirements of the delay-controlled approach.

Network Time Protocol

NTP provides a mechanism to continually request time information from different time servers in the network. The more NTP servers provide information, the higher the accuracy in terms of time offset and frequency shift of the local clock. NTP networks consist of primary and secondary NTP servers. Primary servers are directly connected to a UTC source, while secondary servers are synchronized to UTC via primary servers or other secondary servers and, thus, offer lower time accuracy.

As there is no need for UTC time in the delay-controlled echo cancellation system, the new approach does not exactly correspond to the NTP specification [126]. In an appropriate realization one of the two echo cancellers acts as NTP client and the other one as

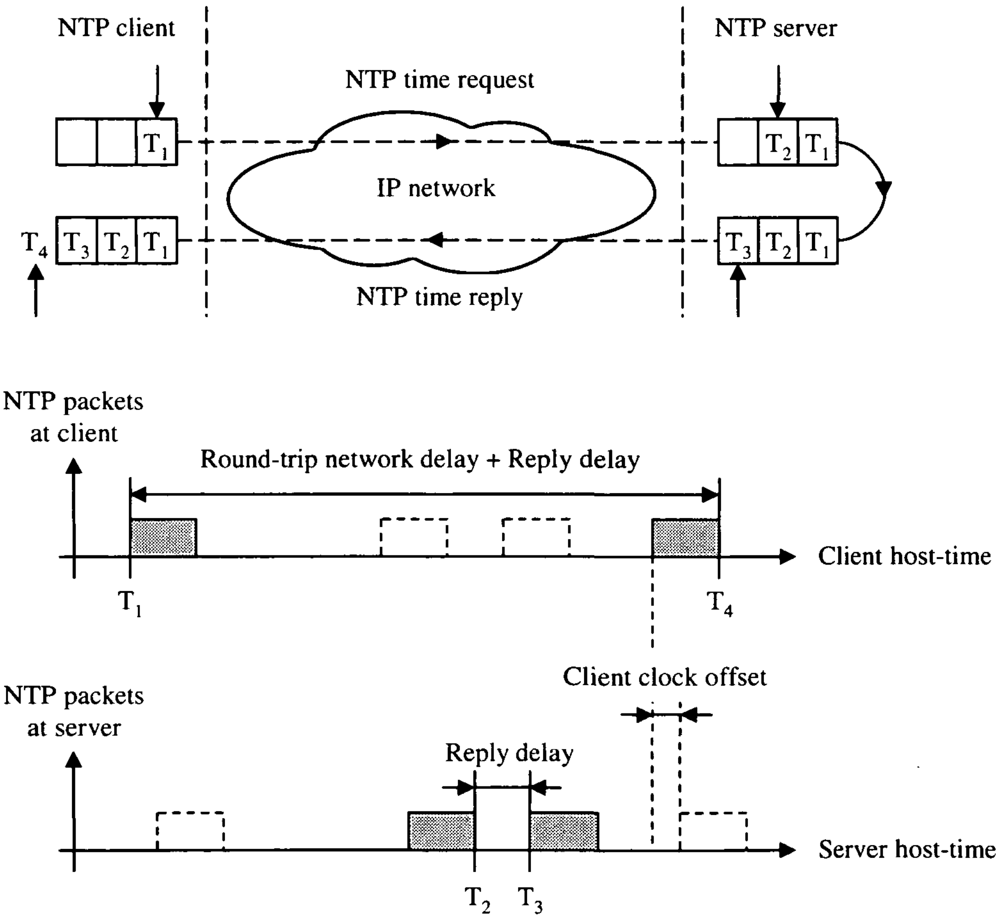

Figure 3.7 – A NTP time request departs the client for the server, which reacts with a NTP time reply. The corresponding timestamps are added to the packet (newly inserted values are marked with an arrow) either prior to the sending process of the first bit or after arrival of the last bit of the NTP packet. The receiving timestamp at the client (T_4) is not part of the NTP packet.

NTP server, whereby the system clock of the latter NTP entity represents the absolute time in the system. Since some environmental conditions like the temperature may have a strong influence on the clock frequency, NTP is also able to correct this parameter of the client clock. The detailed processing of a time query starts with the sending of a time request from the NTP client to the NTP server, which in turn answers with a time reply (Figure 3.7).

The corresponding instants of sending and receiving NTP packets are determined by the timestamps, which are put into the NTP packet just before the transmission of the first bit of the packet and immediately after the receipt of the last bit, respectively. Therefore, the overall round-trip delay includes the network delays in both directions as well as the waiting time at the NTP server. The corrected client time is derived from the present clock value at the receipt of the NTP reply (T_4) by adding an unknown client clock offset:

$$\text{New client time} = T_4 + \text{Client clock offset}. \tag{3.3}$$

The new system time of the local client is also computable from the distant server sending time (T_3) under the assumption of symmetric network conditions:

$$\text{New client time} = T_3 + \frac{\text{Round-trip network delay}}{2}. \tag{3.4}$$

In this context the round-trip network delay summed over both directions excludes the reply delay experienced at the NTP server and amounts to:

$$\text{Round-trip network delay} = T_4 - T_1 - (T_3 - T_2). \tag{3.5}$$

Combining Equs. (3.3), (3.4), and (3.5) results in the desired offset used for the correction of the local time according to Equ. (3.3) [126]:

$$\text{Client clock offset} = \frac{T_2 + T_3}{2} - \frac{T_1 + T_4}{2}. \tag{3.6}$$

The accuracy of synchronization depends on the difference of the delays experienced in both directions between NTP client and NTP server. Therefore, identical serialization and receiving delays at the client and server, respectively, do not change the resulting clock offset. In the case of symmetric paths, i.e. equal network delay values, Equ. (3.6) offers an exact means of local time correction. NTP delivers single-milliseconds synchronization on local-area networks and, at worst, a few tens-of-milliseconds synchronization on wide-area networks. The uncertainties are decreased by using QoS mechanisms such as MPLS (see Section 2.2.1), which reserve a path throughout the network exclusively for NTP messages and, moreover, prioritize the NTP traffic in the IP router queues along that path over other packets. The overall delay between client and server is decreased in this way, and, at the same time, the symmetry of the paths is increased.

GPS receiver

Alternatively, GPS receivers are implementable in both echo cancellers in Figure 3.1, which provide time information accurate to ± 1 μs or better, in order to synchronize to

UTC time. When the time information is obtained directly from a GPS receiver, the operating system and the hardware of the echo control device (which is implemented, for example, on a media gateway) introduces further delays and uncertainties; but with a specific design of those components microsecond-level synchronization in the range of 10s of μs is achievable [43]. Practical realizations use the NTP exclusively for the transfer of the current time values from the GPS to the local delay measurement unit with the intention to avoid interlocks, when two or more systems want to synchronize on the same satellite receiver card. Such a concept has been realized within the IST-project AQUILA [166]. Adversely, GPS cards are coming with high costs; but recent trends show a decreasing tendency.

Exchange of Time Information
Having synchronized both clocks and having evaluated the one-way network delay values for both directions at the corresponding ends of the connection, the time information has to be exchanged between the entities. As the new concept is applicable in corporate networks or in-house communication systems, special IP packets containing the time information are introduced. In order to guarantee exact attenuation control, the packets are transmitted continually at constant time intervals or—after this information has been transmitted once—the exchange is limited to noticeable delay changes.

Signal Processing Delay
Both methodologies for the determination of the round-trip and one-way delay using timestamps are limited to the network delay. In addition to that the time span consumed by other signal processing components like the de-jitter buffers and coding schemes as well as the echo-path delay at the near end also have to be taken into account. This information is directly available on-board or constant values are assumed.

Conclusion on Timestamps
On the whole, timestamps allow the determination of one-way delays for every IP packet without influencing the system behavior. On the one hand, the round-trip network delay is assessable by the means of a modified RTP version. On the other hand, one-way delay values are derived with high accuracy when implementing GPS mechanisms and, moreover, the delay information has to be exchanged between the sending and receiving entity.

3.2.3 Combined Methodology

The concept of an explicit measurement signal is not able to adapt to changing delay conditions. A combination of this methodology with the evaluation of timestamps is realizable: During the call establishment phase the offset between the two clocks is determined by using an explicit signal. Based on this offset, the current one-way delay is calculated from the timestamps.

3.3 Conclusion

The suggested implementation of a delay-controlled non-linear processor for echo cancellers limits echoes below the perception threshold according to the talker echo tolerance curves in [75]. Audible distortions of the near end speech are minimized, when the double talk detection operates unreliably.

Beside the determination of the attenuation along the echo path, the control block also has to measure the echo delay. On the one hand, an explicit measurement signal provides an easy to implement means of synchronization. Disadvantageously, the intrinsic injection may also disturb the network behavior. On the other hand, the one-way and round-trip network delay is derived from the timestamps of the data packets. The determination of one-way values requires synchronized receiver and transmitter clocks, which are either achieved by the deployment of the NTP or by using satellite receiver cards with much higher accuracy. Special IP packets have to be introduced in this case in order to exchange the one-way delay information between both echo control devices. Finally, the evaluation of round-trip network delay values is possible without synchronization, when using a modified version of the RTP, which sends back the received timestamp to its origin.

Chapter 4

Listening Test

> The meaning of the word perception: “In humans, the process whereby sensory stimulation is translated into organized experience. [..] Because the perceptual process is not itself public or directly observable (except to the perceiver himself, whose percepts are given directly in experience), the validity of perceptual theories can be checked only indirectly. That is, predictions derived from theory are compared with appropriate empirical data, quite often through experimental research.”[1]

4.1 Introduction

Due to the insufficient performance of state of the art echo cancellers and the already quite matured means of delay measurement in IP networks, the delay-controlled echo control concept introduced in Section 3 is designed for the deployment in such environments (see Figure 3.1). Especially in the evolving virtual trunking scenario (see Figure 2.1), the hybrids located in the local switches are responsible for significant talker echoes.

The main motivation for the conduction of a third-party listening test is to show the room for improvement of the perception oriented delay-controlled echo canceller in comparison to a center clipping implementation (also referred to as standard, conventional, or traditional approach). As already discussed in Section 2.3.7, there are only proprietary and, thus, no standardized means available for the objective determination of the conversational voice quality. Such algorithms would have to consider not only the listening but also the talking related effects of a conversation such as sidetone, self-masking, and talker echo. Therefore, a subjective methodology has been chosen: The third-party technique, in particular, satisfies the required needs, as it provides high sensitivity for the investigation of different parameter combinations for both concepts. Finally, beside the lack of objective

---

[1] Encyclopædia Britannica.

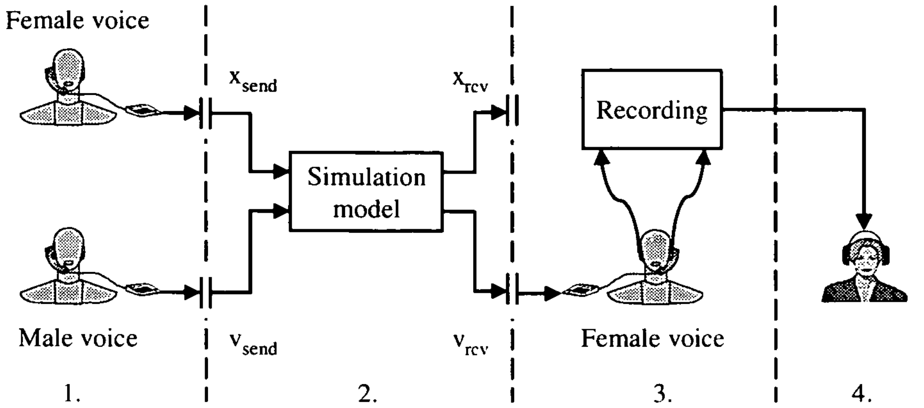

Figure 4.1 – Sequential creation of binaural voice samples and final assessment of the speech sequences by subjects.

methodologies, the end user always decides, whether a product fulfills the quality criteria or not.

Third-party listening tests are mainly meeting criticism, because the participating subjects are assessing voice samples of unknown speakers and, in this way, are not listening to their own voice. In spite of this fact, the results of a comparative evaluation of the third-party technique and the more realistic talking and listening test point out highly correlated MOS values [61] [119]. Moreover, the pure listening test even comes with smaller confidence intervals. The listeners of this test were facing a hands-free situation at different one-way delays in single and double talk situations. As with other auditive investigations, the selection of the source material in this test series has a strong influence on the outcome.

The general setup for the recording and evaluation of voice samples in a third-party test has already been illustrated in Figure 2.24. The special procedure used within this work is based one four sequential steps (Figure 4.1).

Firstly, the signals of a pre-recorded dialogue between a female and male speaker of compact disc (CD) quality[2] have been taken and fed separately into artificial mouths, which are both located in different HATSs. In this context, the corresponding loudspeakers have provided a desired speech level of -4.7 $\mathrm{dB_{Pa}}$[3] at the mouth reference point [P.58]. The sound of the loudspeakers is converted back into electrical signals, when it meets the microphones of each telephone receiver. A standard ISDN phone called “Europa 10” has been utilized for the recordings; it meets the specified requirements according to [156] and [150]. Based on the measured sensitivity at various frequencies, the SLR and RLR values of the “Europa 10” have been calculated by using weighting factors at discrete frequencies [97]:

$$\begin{aligned} \mathrm{SLR} &= 8.38\ \mathrm{dB}, \\ \mathrm{RLR} &= 3.30\ \mathrm{dB}. \end{aligned} \tag{4.1}$$

The voice samples of the male and female speaker, which have been induced at the near

---

[2] CDs store audio signals at a sampling rate of 44.1 kHz and with a resolution of 16 bit.

[3] The unit of the sound pressure level at the mouth reference point ($\mathrm{dB_{Pa}}$) is measured relative to 1 Pascal.

and far end, respectively, have been recorded at the S_0 interface of the ISDN phone as linear signals with 48 kHz sampling rate; the recorded sequences are named v_{send} and x_{send}, respectively. Although the newly introduced concept is designed for mixed analog and digital scenarios, an ISDN phone has been chosen in order to provide well defined and undisturbed input signals for the offline simulation.

In a second step, the various settings of the conventional and the new approach have been simulated offline [26] [27]. The aforementioned voice sequences have been taken as inputs for the system model. Both signals have been sampled down to 8 kHz prior to the simulation, because real-world DSP voice applications are based on such sampling rates in most cases. The simulation outputs the input signal for the far end talker's telephone (v_{rcv}), which should be very close to the original female voice (v_{send}) in order to perceive a high level of conversational voice quality. Besides, the female voice has also been transferred over the IP based telephone network and received as x_{rcv}; but only the female talker echoes occurring at the far end are considered in the third-party listening test and, thus, x_{rcv} is not recorded.

Thirdly, the signal v_{rcv} is injected into the S_0 interface of the female talker's phone. Moreover, the undisturbed reference x_{send} is played back simultaneously according to the original time pattern. The two microphones of the far end HATS record the binaural samples, which include all effects such as handset sidetone and acoustical leakage at both ears (generated by the original female voice x_{send}) as well as female talker echo and male double talk signal (originated by the simulation output v_{rcv}).

Finally, these stereo sequences are judged via free-field equalized headphones by 21 subjects. The listening test is carried out with German mother-tongue subjects, and, thus, speech samples of the same language have been utilized for the auditive test.

4.2 Creation of Voice Samples

4.2.1 Introduction

This section deals with the modeling of the traditional and the new approach of echo cancellation as well as the mathematical description of the test environment including the mixed analog and packet-switched telecommunication network scenario as shown in Figure 4.2. Various settings of both echo cancellers and the network environment are simulated based on these models. A summarization of all parameters—constant as well as variable ones—is finally given in Table 4.2. The corresponding simulation models are listed in Appendix A of this work.

The software package Matlab® (short version of matrix laboratory) has been chosen for this task. Matlab provides an integrated computing environment for technical applications that combines numeric computation, graphics and visualization as well as a high-level programming language. Furthermore, it comes with a variety of toolboxes, which are collections of Matlab® functions that extend the basic environment to solve problems in particular fields of applications; for this work the Simulink® and the DSP Blockset® toolbox have been utilized. Simulink® is designed for modeling and simulating physical and math-

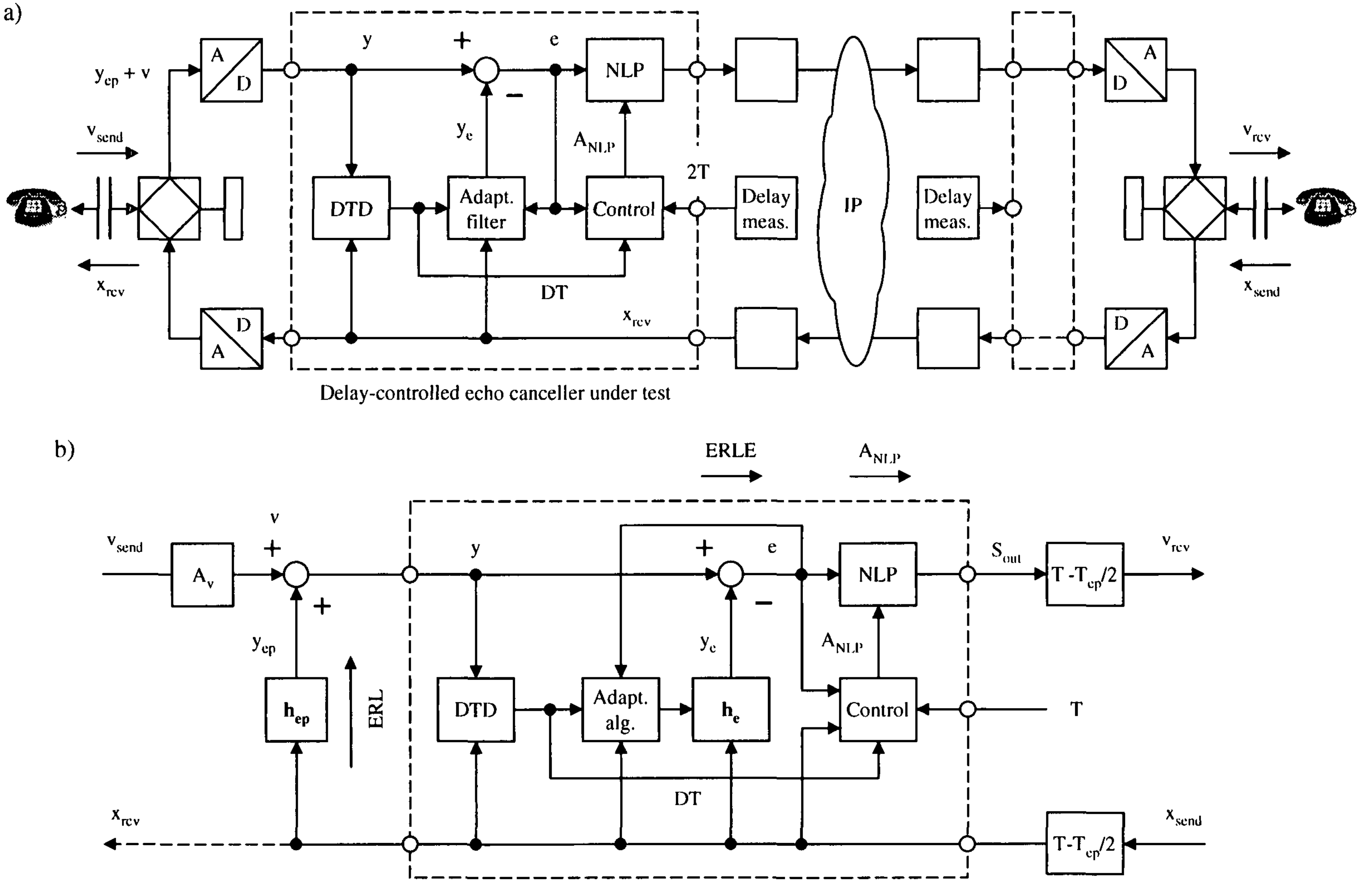

Figure 4.2 – Modeling of the delay-controlled echo canceller deployed in an IP based telephone network, which connects two PSTNs. a) Two analog telephones are assumed to inject the ideal ISDN samples as near and far end signals v_{send} and x_{send}, respectively. The transmitted signal v_{rcv} represents the output of the simulation utilized for the stereo recordings at the far end. b) The corresponding simulation model of the system is equipped with an additional attenuation A_v for the male input voice.

ematical systems and it is often applied with system models of other toolboxes. Beside linear and non-linear models, Simulink® also presents continuous and time-discrete means of simulation; both methodologies have been deployed for the creation of the voice samples. The DSP Blockset® provides a set of functions for realizing digital signal processing components such as adaptive filters or time-frequency transformations.

4.2.2 Echo Canceller

Principally, both, the standard and the delay-controlled model, are made up of the same building blocks with identical parameter settings. They exclusively differ in the design of the NLP control mechanism. On the one hand, the standard approach requires the computation of the NLP clipping level. On the other hand, the delay-controlled design continually adapts the NLP attenuation according to Equ. (3.1).

All the standard components (adaptive filter, DTD, and NLP) and their basic mathematical descriptions have already been introduced in Section 2.3.5. In this section the enhancements necessary for the third-party listening test as well as the special design of both NLP concepts are discussed.

Adaptive Filter

The adaptive filter block is modeled as the widely deployed NLMS algorithm, which represents an adaptive FIR filter using the stochastic gradient methodology (see Section 2.3.5). The column vector $\mathbf{h}_e$[4] of length N = 512 continually estimates the impulse response of the echo-path. The corresponding adaptation algorithm utilizes the estimation error at the output of the subtraction unit (e) and calculates the estimation vector only under single talk conditions (i.e., DT = 0)[5]:

$$\mathbf{h}_e(n+1) = \mathbf{h}_e(n) + \mu \frac{e(n)}{\mathbf{x}_{rcv}^T(n)\mathbf{x}_{rcv}(n)} \mathbf{x}_{rcv}(n) \Bigg|_{DT(n)=0} . \tag{4.2}$$

The system variable $\mathbf{x}_{rcv}$[6] represents the last N input samples on the receiving path and μ corresponds to the adaptation step size, which is always in the range of $0 < \mu < 2$ in order to guarantee converging filter coefficients for every possible combination of input signals. The number of coefficients N gives a direct measure for the window size of the echo canceller by dividing it with the sampling rate of f_s = 8 kHz:

$$\text{Window size} = \frac{N}{f_s} = 64 \text{ ms} . \tag{4.3}$$

The output of the adaptive filter is derived from the estimation vector of Equ. (4.2):

---

[4] The time-discrete notation of a system variable such as $\mathbf{h}_e(n)$ is only used in numbered equations, while the spelling $\mathbf{h}_e$ is utilized within the text.

[5] The input parameters are sampled at the discrete time instant n, while the outputs and all in between computations are denoted at n+1.

[6] The bold spelling of variables indicates vectors, while plain fonts are used for scalars; both syntaxes are possible for the same variable.

$$y_e(n+1) = \mathbf{h}_e^T(n+1)\mathbf{x}_{rcv}(n). \tag{4.4}$$

In order to show the system performance independently of the adaptive filter, the NLMS algorithm is in some cases replaced by a simple gain block, which provides the linear amplification K_{NLMS}. The estimated system response of the echo-path is derived artificially, as the signal y_{ep} is not assessable in real-world scenarios, by calculating

$$y_e(n+1) = K_{NLMS}\, y_{ep}(n), \tag{4.5}$$

which results in a constant filter attenuation in terms of ERLE.

DTD Algorithm

The DTD is modeled as enhanced version of the level-based Geigel algorithm, which has been introduced in Equ. (2.10). The output is set to DT = 1, when

$$\begin{aligned} |y(n)| &> G_{DTD} \max|\mathbf{x}_{rcv}(n)| \text{ and} \\ |y(n)| &> L_{min}. \end{aligned} \tag{4.6}$$

In any other case DT is set to zero. As often implemented in practice, the number of observed input samples, which corresponds to the length of vector $\mathbf{x}_{rcv}$, equals the number of FIR filter taps used within this work:

$$N_{DTD} = N = 512. \tag{4.7}$$

The parameter L_{min} in Equ. (4.6) has been introduced in order to react only in the case of significant amount of near end speech; a constant value of

$$L_{min} = 2.10^{-3} \tag{4.8}$$

meets this requirement. The parameter G_{DTD} depends on the expected echo-path attenuation; a rough estimate is given by the following equation:

$$G_{DTD} \approx 10^{-\frac{ERL}{20}}. \tag{4.9}$$

A small value for G_{DTD} results in undetected double talk periods and, consequently, non-linear distortions are occurring in the near end signal in the case of a center clipper, while, on the other hand, the new approach just attenuates the near end signal without disturbing it. When choosing a high value for G_{DTD}, the DTD decides on double talk in some far end single talk situations, which has the consequence of non-cancelled talker echoes at the far end side for both implementations. A setting of

$$G_{DTD} = \frac{1}{\sqrt{2}} \tag{4.10}$$

turned out to provide reliable detection behavior.

Moreover, the DTD always holds its output of DT = 1 for a minimum period of time; the corresponding holding time has been chosen to:

$$T_{hold} = 200 \text{ ms}. \tag{4.11}$$

The relative level parameter as well as the holding time according to Equs. (4.10) and (4.11), respectively, have been optimized in a separate test session within the listening test.

NLP Characteristic and Control Mechanism

As mentioned before, the listening test aims at a comparison of the standard and the delay-controlled approach, which exclusively differ in the implementation of the NLP control mechanism under far end single talk conditions.

Both NLP concepts perform identically in transferring the input signals unchanged to the output, when the DTD decides on double talk:

$$S_{out}(n+1) = e(n)\big|_{DT(n)=1}. \tag{4.12}$$

The corresponding control mechanisms of the conventional and the new concept deliver $L_{clip} = 0$ and $A_{NLP} = 0$ dB, respectively, in such cases. The NLPs perform as linear characteristics with 0 dB attenuation. The real double talk mode considering the masking of echo signals by the near end signal has not been taken into account within this listening test.

The following describes the different behaviors in single talk situations. On the one side, the standard NLP functions as a center clipper with adaptive and symmetric switching levels ($-L_{clip}$ and L_{clip}) according to the left-hand diagram of Figure 2.22. Input levels above the symmetric clipping levels are treated as if the system would be in the double talk mode as described by Equ. (4.12). The NLP thresholds $-L_{clip}$ and L_{clip} rely on a short-time estimate of the magnitude of the residual echo and they are measured only under far end single talk conditions:

$$L_{clip}(n+1) = K_{NLP} \max |\mathbf{e}(n)|_{DT(n)=0}. \tag{4.13}$$

The length of the system vector **e** corresponds to the NLP time window of 10 ms; the corresponding number of observed samples is, consequently, set to

$$N_{NLP} = 80. \tag{4.14}$$

The assurance factor K_{NLP} is always slightly larger than one

$$K_{NLP} = 1.01, \tag{4.15}$$

and it guarantees the suppression of residual echoes. The shift register implemented for the calculation of Equ. (4.13) is not flushed when DT = 1; the collection of samples just stops under this condition and carries on when DT = 0.

On the other side, the NLP of the delay-controlled echo canceller comes with a variable and linear attenuation, which is described by the parameter A_{NLP} in dB according to Equ. (2.1):

$$A_{NLP}(n+1) = TELR[T(n)] - [ERL(n+1) + ERLE(n+1) + SLR + RLR]. \tag{4.16}$$

The resulting output is calculated as follows:

$$S_{out}(n+1) = 10^{-\frac{A_{NLP}(n+1)}{20}} e(n)\Big|_{DT(n)=0} . \qquad (4.17)$$

The new concept utilizes a linear characteristic (see "far end single talk" curve in Figure 3.3) without considering the masking in real double talk periods. The control mechanism determines the required NLP attenuation according to Equ. (4.16). The accuracy relies on the exactness of the determination of the echo delay and echo attenuation. The simulation of every single voice stream transferred across the telephone system is based on a constant mean one-way delay T; several single calculations have been carried out in order to investigate the relevant time span of the system. The exact filter attenuation of the echo signals results from the system response of the echo-path (y_{ep}) and the residual echo (y_{ep} - y_e) [1]:

$$\mathrm{ERLE}(n+1) = 10\lg \frac{\mathrm{LPF}\left\{\left[y_{ep}(n)\right]^2\right\}}{\mathrm{LPF}\left\{\left[y_{ep}(n) - y_e(n)\right]^2\right\}}\Bigg|_{DT(n)=0} , \qquad (4.18)$$

whereby Equ. (4.18) provides an artificial measure, since the signal y_{ep} is not accessible in real-world scenarios. The function LPF{·} defines a lowpass filter operation of first order, which has the following appearance in the time discrete domain

$$\mathrm{LPF}(z) = \frac{\frac{z}{T_{LPF}}}{z - e^{-\frac{1}{f_s T_{LPF}}}} . \qquad (4.19)$$

The time parameter T_{LPF} represents a compromise between smooth and easy control and very dynamic tracking of changes:

$$T_{LPF} = \frac{1}{8} s. \qquad (4.20)$$

As the output of the hybrid (y_{ep}) in Equ. (4.18) is not measurable, the adaptive filter performance is approximated under far end single talk conditions by assuming that $y_{ep} \approx y$ and, as a consequence of this, $y_{ep} - y_e \approx e$:

$$\mathrm{ERLE}(n+1) \approx 10\lg \frac{\mathrm{LPF}\left\{[y(n)]^2\right\}}{\mathrm{LPF}\left\{[e(n)]^2\right\}}\Bigg|_{DT(n)=0} , \qquad (4.21)$$

whereby the system memories of the lowpass filter implementation are not reset under double talk conditions. Alternatively, a practical implementation also directly assesses the whole echo attenuation from the receiving path to the egress of the sending path:

$$(\mathrm{ERL} + \mathrm{ERLE})(n+1) \approx 10\lg \frac{\mathrm{LPF}\left\{[x_{rcv}(n)]^2\right\}}{\mathrm{LPF}\left\{[e(n)]^2\right\}}\Bigg|_{DT(n)=0} . \qquad (4.22)$$

The positive attenuation of the echo-path is directly computable from the coefficients of the adaptive filter:

$$\mathrm{ERL}(n+1) = -10\lg\|\mathbf{h}_e(n)\|_2^2 . \tag{4.23}$$

As shown later on in this chapter, the combined results of Equs. (4.21) and (4.23) provide a reliable estimate of the desired echo attenuation from the input of the echo canceller to the filter output.

Moreover, assumptions on the far end terminal characteristics are made according to Table 3.1(i.e., SLR + RLR = 10 dB) in order to derive the required NLP attenuation.

4.2.3 Test Environment

Echo-Path

The model for the hybrid circuit located at the cancelled end is taken from the corresponding ITU-T Recommendation [80]. The so-called "echo path model 1" is characterized by short dispersion; the parameters are the echo-path delay T_{ep} and the echo-path attenuation

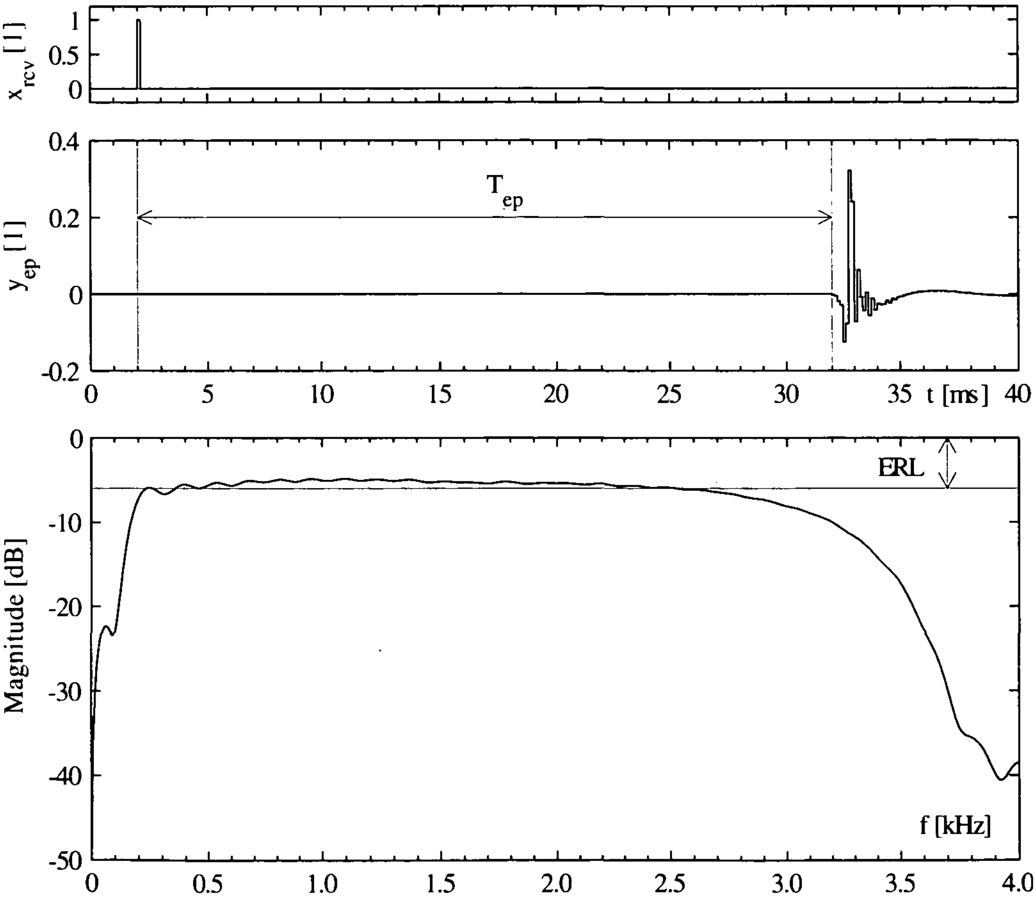

Figure 4.3 – Impulse response and magnitude transfer characteristic of the echo-path model with T_{ep} = 30 ms and ERL = 6 dB.

ERL. The impulse and the magnitude response are illustrated in Figure 4.3.

Attenuation of the Double Talk Signal

The task of the active gain control entity is to adjust the speech levels in the system to standardized values. During a conversation the voice levels are varying due to changing talker behaviors or decreasing when the signals have to traverse analog networks over long distances. The block with parameter A_v immediately placed after the input of the near end signal in Figure 4.2 b) represents this attenuation of the near end signal. As already mentioned in Section 2.1.3, the long-term average speech level should ideally be around -18 dB_{m0}.

Telephone

The "Europa 10" provides the input of the source speech and it is deployed for the stereo recordings of the voice streams, as well. Since this ISDN phone comes with high sidetones (STMR = 18.4 dB), the resulting masking of the received voice samples makes some degradation unnoticeable. Therefore, beside the original setting of the used telephone (also called silent telephone) a second, artificial one, has been conceived (named loud telephone). The characteristics of the loud phone, which are listed in Table 4.1, are identical with the Europa 10; the only difference is the decreased RLR of 0.3 dB, which has the consequence of louder (plus 3 dB) and, thus, more audible, receiving signals at the listener's ear. The resulting RLR is still within the recommended range according to ANSI and ETSI (see Table 3.1). The loud telephone is realized in the recording procedure by reducing the original female voice by 3 dB and leaving, at the same time, the telephone input signal unchanged. Consequently, the corresponding output pressure of the female talker has been measured 3 dB_{Pa} below the standardized level of -4.7 dB_{Pa}. Afterwards in the voting session, the level is increased by 3 dB again in order to provide correct sidetones at both ears of the HATS. The resulting loudness ratings as well as the "SLR + RLR" values for both types of telephones are listed in Table 4.1. The artificial phone comes even closer to the assumed value of RLR + SLR = 10 dB, which is assumed by the NLP control algorithm of the new approach.

IP Network

The different one-way talker echo delays T are experienced across the mixed analog and IP based telephone network. In order to show the room for improvement of the new system, the model of the interface to and the IP part itself is restricted to discrete and constant end-

	SLR [dB]	RLR [dB]	SLR + RLR [dB]
Silent telephone	8.38	3.3	11.68
Loud telephone	8.38	0.3	8.68

Table 4.1 – Loudness ratings for the original, silent Europa 10 telephone and for the loud, modified phone.

to-end delay values, which are equal in both directions:

$$T' = T - \frac{T_{ep}}{2}. \tag{4.24}$$

Therefore, the overall time span experienced by the signals from the talker's microphone to the talker's loudspeaker equals the desired value of 2T, while the received and wanted signals v_{rcv} and x_{rcv} undergo T' of Equ. (4.24). Besides, the discrete values listed in Table 4.2 are chosen logarithmically, as it has also been done for the standardized talker echo tolerance curves (see Figure 2.16), in order to cover a broad time range. As a consequence of the varying time patterns, separate inputs (v_{send} and x_{send}) for every single one-way delay value have been provided by inserting the required delay values according to Equ. (4.24) between the talking spurts. Furthermore, all voice packets are arriving at the addressee; i.e., the packet loss of the IP network is set to zero.

The time pattern of the communication sequences changes with different delay values as depicted in Figure 4.4. The overall length of the male and the female voice stream are enlarged by $4T'$ and $3T'$, respectively, due to the end-to-end delay inserted between the talking spurts. When the participants have finished their talking sequences, the speech signals traverse the telephone network and reach the destination after T'. Then, the other talker needs consideration time, until another message is sent back to the origin; this procedure within a conversation is called *turn-taking*. The only exception is the second and, at the same time, last talking spurt of the female voice, since it is not influenced by the male reply delay. Consequently, when varying the one-way delays, the male voice masks different parts of the female speech during double talk.

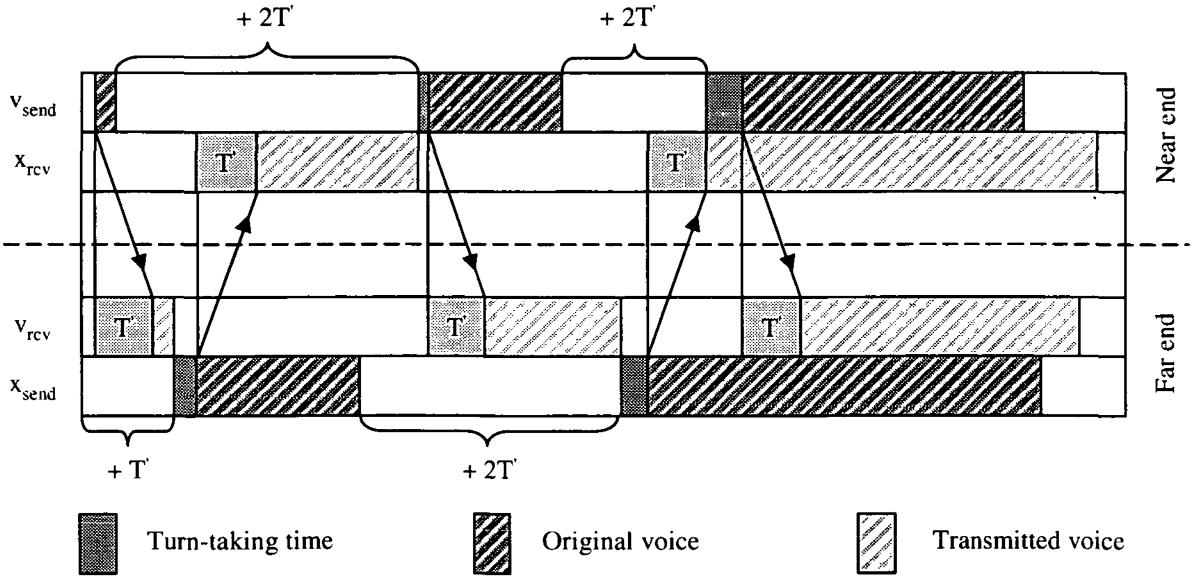

Figure 4.4 – Change of the communication time pattern due to different one-way delays across the telephone system. The original voice sequences need the time T' to arrive at the receiving side and after a short waiting time the reply is returned.

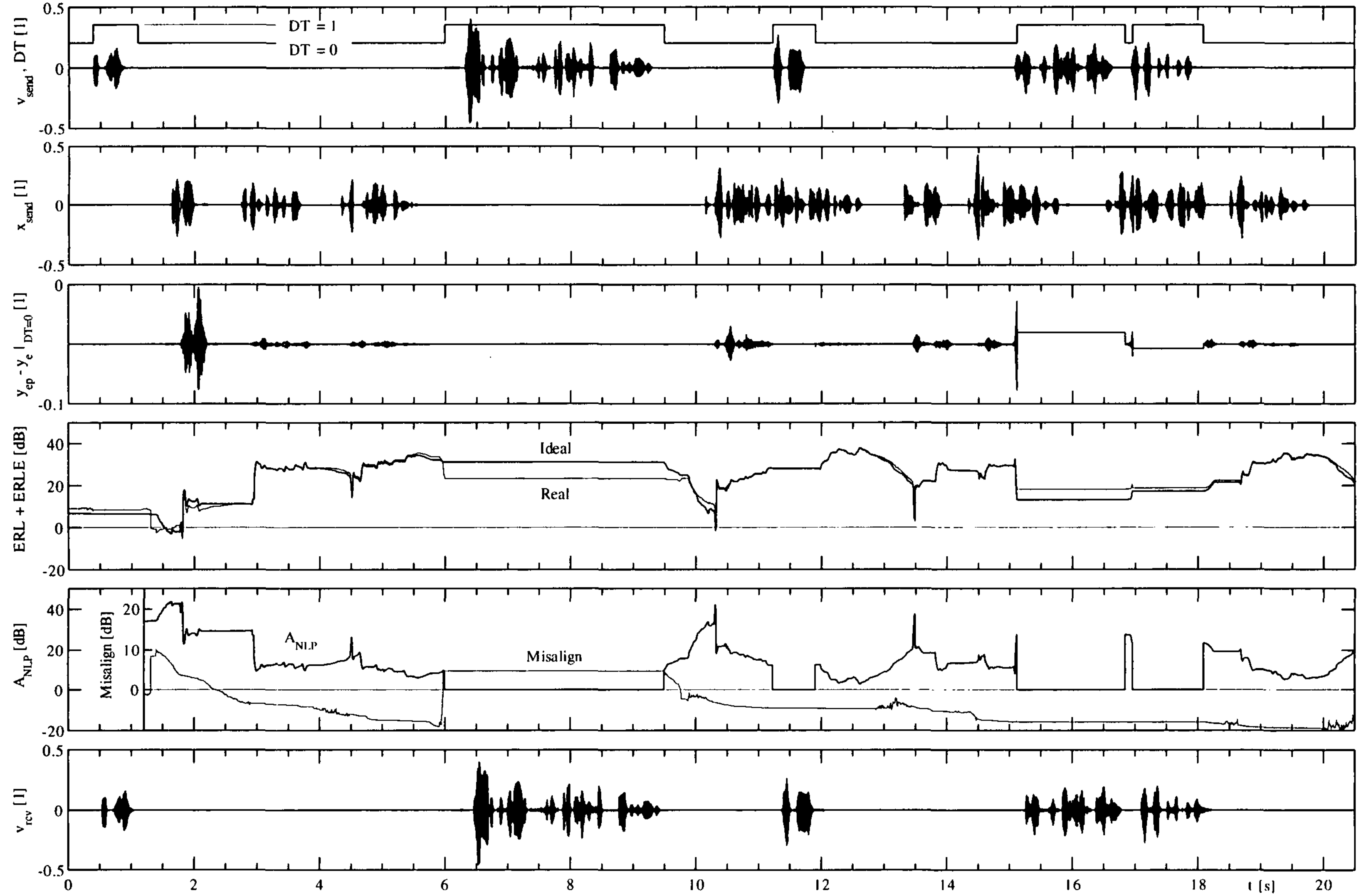

Figure 4.5 – The delay-controlled echo canceller faces a telephone network with T = 160 ms. From the top to the bottom diagram: Firstly, injected near end, male voice v_{send} and the corresponding DT signal. Secondly, far end, female voice x_{send} applied in the opposite direction. Thirdly, modified residual echo y_{ep} - y_e updated only under single talk conditions. Fourthly, overall echo performance "ERL + ERLE" determined ideally and with real-world methodologies. Fifthly, resulting NLP attenuation A_{NLP} based on the ideal filter performance assuming ERL = 6 dB and SLR + RLR = 10 dB, while TELR = 50.6 dB; the convergence state of the adaptive filter coefficients (misalign). Sixthly, transferred near end speech v_{rcv} of the new approach.

4.2.4 Overall Simulation Results

The special dialogue used for the listening test is represented by the male and female input samples, which are referred to as v_{send} and x_{send}, respectively; both are plotted in the top two diagrams of Figure 4.5. As often implemented in practice, the simulation processes both streams at a sampling frequency of 8 kHz instead of the original recording rate of 48 kHz. The resolution used for the simulation remains unchanged compared to the pre-recorded source material; it amounts to 16 bit. The average active speech level of the samples equals -17.8 dB_{m0} and is, thus, very close to the recommended level of -18 dB_{m0} (see Section 2.1.3). The conversation starts with a short greeting sequence of the male speaker, followed by a single talk period of the female voice of about 4 s, which allows the adaptive filter to converge. Then, the male participant sends a statement of 3.5 s length. In the second half the female voice is interrupted twice by two double talk inputs of the male user, which last for 0.6 s and 1.1 s.

Both input signals in Figure 4.5 undergo a one-way delay of $T' = 145$ ms. Consequently, the one-way delay of the talker echo amounts to T = 160 ms according to Equ. (4.24), when assuming an echo-path delay of $T_{ep} = 30$ ms (see Table 4.2). Moreover, the male voice is not attenuated (i.e., $A_v = 0$), which results in $v = v_{send}$.

Additionally, the output of the DTD block (DT) is drawn with the male voice v_{send} in Figure 4.5. When the DTD indicates double talk (i.e., DT = 1), the near-end echo canceller shows the following behavior. Firstly, as already mentioned, the adaptation process of the filter algorithm is stopped. The parameter ERLE represents the filter performance in terms of the attenuation of talker echo signals. Alternatively, the filter convergence state is represented independently from the voice signals and, thus, more objective, when considering the deviation of the estimation vector:

$$\text{Misalign}(n+1) = 10\lg\left\|\frac{h_e(n) - h_{ep}(n)}{h_{ep}(n)}\right\|_2^2 . \tag{4.25}$$

The "ERL + ERLE" curve is shown in the fourth subplot, while the "misalign" graph is illustrated in the fifth diagram of Figure 4.5. Both plots hold their value when DT = 1. The maximum convergence state of the "misalign" measure equals -10 dB and is reached at about t = 6 s and t = 20 s, where the two longest far end single talk periods end. The worst "misalign" value of about 5 dB is obtained at t = 6 s, where the DTD needs some reaction time to detect the beginning of the male sequence. During this reaction time the adaptive filter diverges by about 15 dB in terms of the "misalign" measure. Secondly, both NLP implementations are switched into transparent mode under double talk conditions. The corresponding value for the NLP attenuation of $A_{NLP} = 0$ dB is illustrated in the same plot as the "misalign" measure. Moreover, the value of the artificial

$$\text{Residual echo}(n+1) = \left[y_{ep}(n) - y_e(n)\right]\Big|_{DT(n)=0} , \tag{4.26}$$

which is used for the calculation of the exact echo attenuation according to Equ. (4.18), is held at its current value during double talk periods. Throughout the first far end single talk phase, which lasts from t = 1.7 s to t = 5.6 s, the remaining echo level in the third diagram of Figure 4.5 is continually decreased, because of the converging adaptive filter.

The delay-controlled NLP attenuation (A_{NLP}) is calculated exclusively under single talk conditions according to Equ. (4.16). Ideally, the graph in Figure 4.5 is based on the ERLE computation of Equ. (4.18), which has been utilized for the creation of all voice samples.

Moreover, the control mechanism artificially receives the hybrid attenuation of ERL = 6 dB as input and it assumes

$$\mathrm{SLR} + \mathrm{RLR} = 10\,\mathrm{dB}\,. \tag{4.27}$$

The A_{NLP} graph supplements the ideal "ERL + ERLE" curve in order to fulfill Equ. (4.16). The required TELR value for T = 160 ms equals 50.6 dB. Additionally, the real sum of ERL and ERLE, which is based on Equs. (4.21) and (4.23), respectively, is drawn with the ideal measure in Figure 4.5.

The last plot in Figure 4.5 shows the delayed output of the sending path of the delay-controlled echo canceller (v_{rcv}), which is recorded for the input for the far end telephone. Differences to the undisturbed reference v_{send} in terms of talker echoes and clipped speech segments are mainly occurring in the two double talk periods of the communication. The distortions are not observable in Figure 4.5 due to the time scale of the illustration and the high overall echo attenuation TELR, which results in almost completely suppressed signals in single talk periods. As another consequence of the latter reason, the simulation output of the standard approach for T = 160 ms is not distinguishable from the delay-controlled one.

In order to show the delay-controlled improvement more clearly, a detailed illustration for T = 15 ms is given in Figure 4.6. The system response y_{ep} of the echo-path represents, more or less, a delayed and attenuated version of the received female voice x_{rcv}. The corresponding parameters are the echo-path delay T_{ep} = 30 ms and the echo-path attenuation ERL = 6 dB (also see Figure 4.3). Moreover, the standard implementation completely suppresses the residual echo under far end single talk conditions (i.e., parts of the speech are clipped), while the new approach attenuates the output of the subtraction unit according to the NLP attenuation (A_{NLP}). Both concepts transfer the NLP input unchanged to the output for DT = 1.

All in all, the perception-oriented concept offers a higher listening quality of the transferred male voice, since significant parts of the desired male voice are transferred (with lower amplitude) and not cut away as with the conventional NLP realization.

All simulations have been started with an initialization sequence consisting of a short noise sequence (of about 50 ms) in the receiving path of the echo canceller. This part is not illustrated in Figure 4.5. Finally, the block diagrams of the simulations and the corresponding parameters are found in Appendix A.

4.2.5 Test Conditions

Table 4.2 summarizes all parameter settings used for the simulation in order to generate the required voice samples, which are applied either for both types of echo cancellers or for the environmental telephone system. The standard values correspond to the default values for

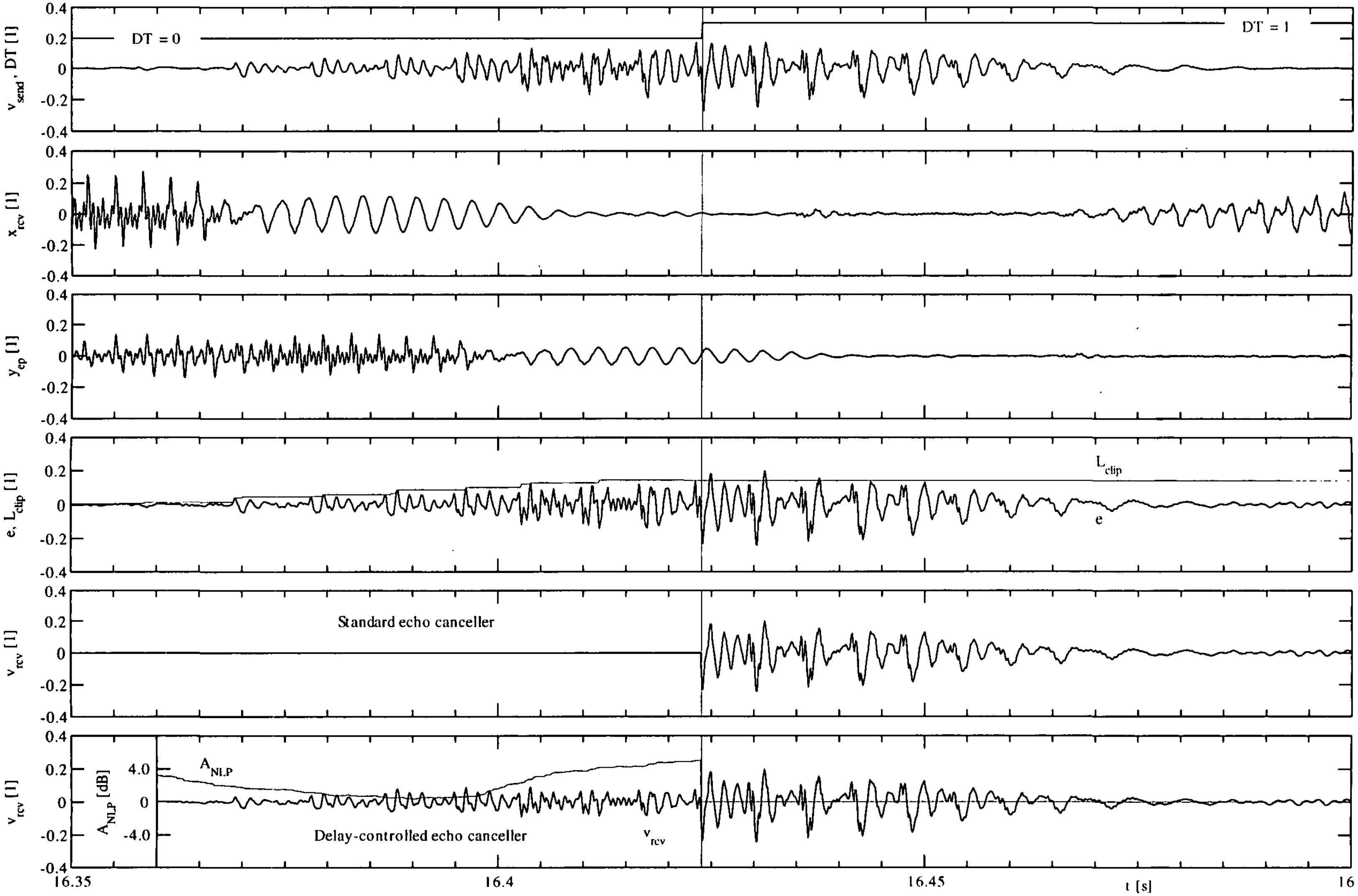

Figure 4.6 – Front end clipping and attenuated transmission at the beginning of a male voice sequence for the standard and the delay-controlled echo canceller, respectively, for T = 15 ms. From the top to the bottom diagram: Firstly, male voice v_{send} and the corresponding DT signal. Secondly, received female voice x_{rcv}. Thirdly, delayed and attenuated output y_{ep} of the echo-path. Fourthly, residual echo e at the output of the subtraction unit and the clipping level L_{clip} for the conventional NLP. Fifthly, output of the system v_{rcv} for the standard echo canceller, which completely suppresses male voice samples of 60 ms length. Sixthly, signal v_{rcv} in the delay-controlled case, which equals the attenuated (A_{NLP}) version of the e signal.

			Standard values	Variants
Echo canceller	NLMS filter		$N = 512$	without filter (→ $ERLE \equiv 0$ dB)
			$\mu = 0.2$	
			y_e → Equ. (4.2)	
	DTD		$G_{DTD} = 0.707$	$G_{DTD} = 0.625, 1$
			$T_{hold} = 200$ ms	$T_{hold} = 30, 100, 400$ ms
			$N_{DTD} = 512$	–
			Decision → Equ. (4.6)	–
	NLP control	Standard	$K_{NLP} = 1.01$	–
			$N_{NLP} = 80$	–
		Delay-controlled	$T_{LPF} = 0.125$ s	–
			SLR = 7 dB	–
			RLR = 3 dB	–
			ERL → Equ. (4.23)	ERL + ERLE → Equ. (4.22), real measurement
			ERLE → Equ. (4.21), ideal measurement	
			A_{NLP} → Equ. (2.1)	–
Test environment	Input signals		$f_s = 8$ kHz	–
			16 bit resolution	–
	Echo-path		ERL = 6 dB	ERL → ∞, for double talk references
			$T_{ep} = 30$ ms	–
	Attenuation of near end signal		$A_v = -6$ dB	$A_v = 0$ dB, A_v → ∞, for echo references
	IP network		T = 15, 30, 50, 90, 160, 300 ms	
	Telephones (Table 4.1)		Loud	Silent

Table 4.2 – Standard parameter settings and variants for the delay-controlled and the standard echo canceller as well as for the test environment.

the various parameters, while the variants represents alternatives used for special settings such as optimization tests or references.

The third-party listening test evaluates three different attributes of the conversation. Every aspect corresponds to one of the four test sessions listed in Table 4.3.

Test session	Voted attribute
A	Overall quality of the male voice
B	Drop-outs and speech gaps in the male voice
C1 and C2	Echo-disturbance caused by the female voice

Table 4.3 – Test sessions of the third-party listening test and the corresponding attributes, which have been evaluated by the subjects.

Tables 4.4 and 4.5 give a detailed overview of the parameter combinations for every test session in order to document the settings of every single vote.

In the first category, the so-called "standard double talk" test in Table 4.4, all three attributes are assessed and the highest number of votes is given. In order to clearly point out the degradations introduced by the NLP, one parameter set even excludes the adaptive filter. Additionally, the decision process of the DTD is made worse and, consequently, the error potential is increased, when the near end voice is attenuated by A_v = -6 dB. Such conditions—divergent filter coefficients, and low voice levels—occur quite often in real-world scenarios. They are applied in this test to clearly point out the room for improvement of the new approach. Moreover, the deviation of the results of the loud telephone from the silent one are investigated. The last column of Table 4.4 references to the diagram presenting the corresponding results.

Secondly, in the "DTD optimization" test of Table 4.4, the two most important parameters of the DTD (G_{DTD} and T_{hold}) are varied in order to find the optimal settings. The voice samples are generated, again, without an adaptive filter and with reduced near end level.

Finally, there are two kinds of references assessed within the third-party listening test, which are discussed in the following section.

4.2.6 References

The individual settings for the simulation of both reference types, which are either used for the standard double talk test or for the evaluation of echoes under far end single talk conditions, are listed in Table 4.5. The corresponding block diagrams of the simulation models are listed in Appendix A.

On the one hand, the "double talk references" have been recorded under ideal conditions and, in this way, provide the best quality for every distinct parameter combination. Ideally, an echo-free environment with infinite ERL, as shown in Figure 4.7, has been created. The echo canceller is not required in such a network and has, therefore, been excluded from the model. Under these environmental conditions, the female talker receives totally undisturbed voice samples from the male opponent. The reference samples exclusively differ in the received voice level, the end-to-end delays, and the loudness of the listener's telephone. Since those references are used within the standard double talk test, they are assessed in terms of all three attributes listed in Table 4.3.

	Echo canceller				Test environment			Number of					Figure
	DTD		Adaptive filter	NLP Control	A_v [dB]	Tele-phone	T [ms]	Files	Votes in test session				
	G_{DTD} [1]	T_{hold} [ms]							A	B	C1	C2	
Standard double talk	0.707	200	with	standard, delay-controlled	0	Loud	15, 30, 50, 90, 160, 300	12	**12**	**12**	**12**	0	(5.1)
			without		-6			12	**12**	**12**	**12**	0	(5.2)
			with					12	**12**	**12**	**12**	0	(5.3)
						Silent	15, 90, 300	6	**6**	0	0	0	(5.4)
DTD optimization	0.707	30	without	standard, delay-controlled	-6	Loud	15, 90, 300	6	**6**	0	0	0	(5.5)
		100						6	**6**	0	0	0	
		400						6	**6**	0	0	0	
	0.625	200						6	**6**	0	0	0	(5.6)
	1							6	**6**	0	0	0	
							Total	72	72	36	36	0	

Table 4.4 – Echo canceller and test environment parameter variations for the standard double talk and the DTD optimization test. Resulting number of files and the corresponding number of samples for the different test series. The delay-controlled NLP control unit assumes SLR + RLR = 10 dB.

			Near or far end single talk	A_v [dB]	ERL [dB]	Tele-phone	T [ms]	TELR objective [dB]	Number of					Figure
									Files	Votes in test session				
										A Note 2	B	C1	C2	
Double talk references			Near end single talk	0	∞	Loud	30	∞	1	**1**	**1**	**1**	0	(5.1)
				-6			300		1	**1**	**1**	**1**	0	(5.2), (5.3)
						Silent	15		1	**1**	0	0	0	(5.4)
Echo references	Talker echo tolerance curves	1% Note 1	Far end single talk	∞	TELR - (SLR + RLR)	Loud	15	28.2	2	0	0	0	**2**	(5.7)
							300	54.6	2	0	0	0	**2**	
		10%					15	22.2	1	0	0	0	**1**	
							300	48.6	1	0	0	0	**1**	
	Variable TELR						15	10.2, 16.2, 34.2, 40.2	4	0	0	0	**4**	(5.8)
							300	36.6, 42.6, 60.6	3	0	0	0	**3**	
								Total	16	3	2	2	13	

Table 4.5 – Parameter settings for the double talk references and for the echo references test. The latter is divided in sessions for the scan of the "acceptable" (1%) and of the "limiting case" (10%) talker echo tolerance curves as well as for a vertical scan (variable TELR) at the corresponding time values. The echo cancellation is not applied in these tests. The calculation of the ERL values for the echo references test are based on SLR + RLR = 10 dB. Note 1: Beside this standard values, the exact values of the loud telephone (SLR + RLR = 8.68 dB) are also assumed for the ERL computation in order to exactly meet the "acceptable" curve. Note 2: The standard double talk test includes three reference votes; two of these votes are also used as references for the DTD optimization diagrams.

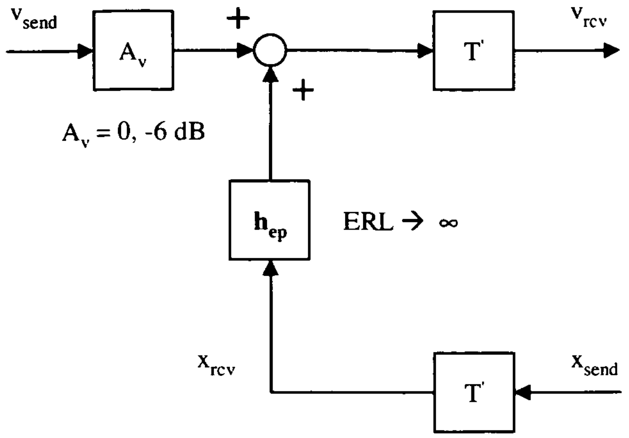

Figure 4.7 – Block diagram for the creation of the double talk reference samples in a completely echo-free telephone environment.

On the other hand, in order to give a reference to the talker echo tolerance curves, the "echo references" in Table 4.5 exclusively induce the female voice samples, while the male talker is completely suppressed (Figure 4.8). The hybrid attenuates the incoming female signals according to the desired TELR value, which depends on the mean one-way talker echo delay. In this test series the subjects evaluate the annoyance of talker echoes caused by the female voice in two cases. Firstly, values from the "acceptable" and "limiting case" curve are taken at T = 15 ms and T = 300 ms (Figure 4.9) in order to follow the "talker echo tolerance curves" of Table 4.5. Beside the standard assumption on loudness ratings of Equ. (4.27), the NLP control algorithm of the delay-controlled approach also provides the exact values of the deployed telephone with the intention to show the error introduced by the standard values. Secondly, a vertical scan at the aforementioned delay values is carried out ("variable TELR").

4.3 Assessment

This section presents important facts concerning the subjects before and during the test as well as the assessed attributes of the conversation between the male and female speaker.

4.3.1 Test Setup

The digital output of a commercially available CD player has been connected to an equalizer, which compensates for two aspects.

Firstly, during the recording process, the simulated voice samples pass the telephone and its receiver, and then the resulting sound wave traverses the ear trumpets of the HATS, before the microphone of the receiver's ear is finally reached. Simultaneously, the female talker's voice has been replayed by the loudspeaker in the HATS's mouth in order to provide for the sidetones, which is perceived by both artificial ears. The transfer function from the HATS's microphones to the ear reference point, which is right in front of the ears, is compensated; this procedure is referred to as *free-field equalization.*

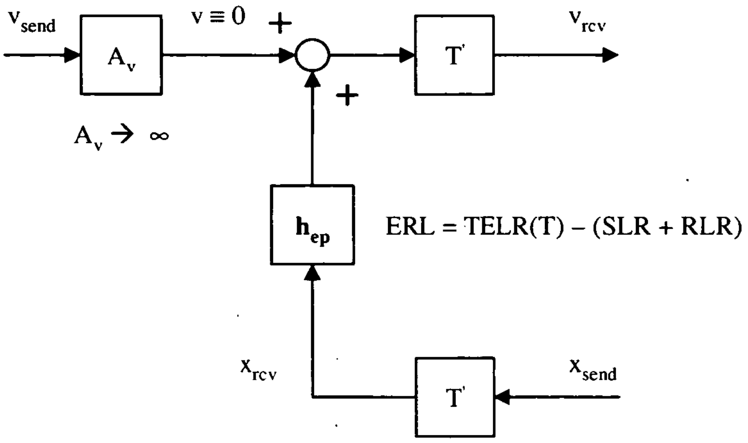

Figure 4.8 – System model used for the echo reference samples, which are assessed under far end single talk conditions.

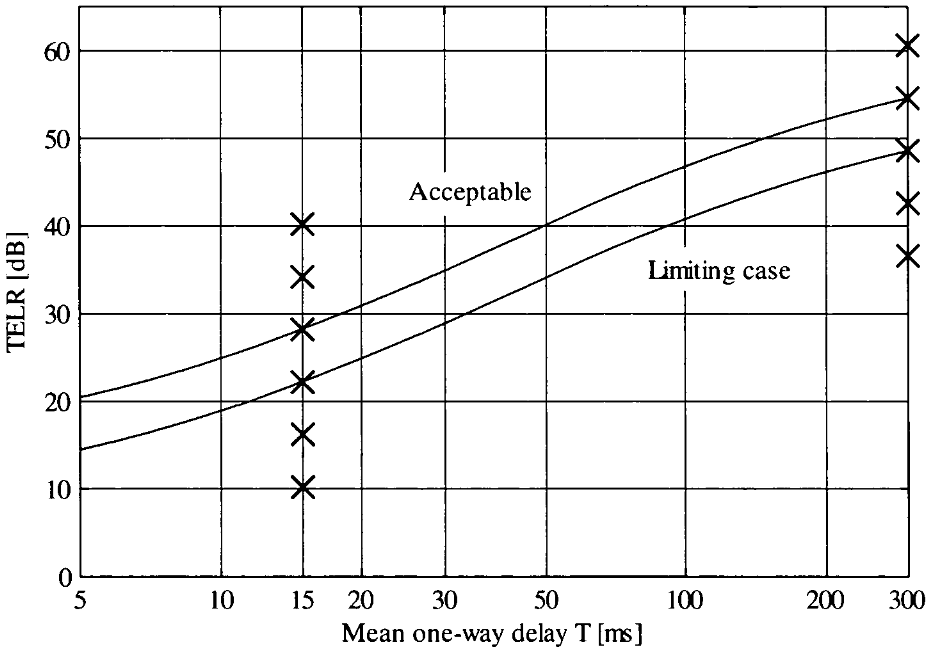

Figure 4.9 – Scan of the talker echo tolerance curves at T = 15 ms and T = 300 ms.

Secondly, the subjects utilize headphones for listening to the stereo voice samples. The transfer characteristic from the electric interface to the ear reference point via those headphones is also compensated.

A linear amplifier is placed between the equalizer and the headphones in order to provide the correct levels with adequate power.

4.3.2 Selection of Subjects

Basically, the corresponding ITU-T Recommendation for subjective testing of network echo cancellers [107] does not give guidelines for the selection and the number of subjects. In order to provide reliable results at least 20 persons should carry out the test; but even a smaller number (down to about 12) may be sufficient, when the confidence intervals are small enough. The third-party listening test has been conducted by 17 male and 4 female subjects and the average age has been 29 years.

4.3.3 Time Schedule

The whole subjective test consists of four test sessions (Table 4.4), which are defined as sequences of samples presenting the same part of the conversation and, moreover, one common aspect is evaluated. A session should contain the whole range of assessable qualities, e.g., votes from very good to very bad in the case of an ACR test. Otherwise, the subjects may change their subjective scale acquired during a session in order to cover the whole range with their single votes. Before starting a test session, a short training sequence consisting of three to seven samples is presented (Table 4.4). It covers a broad quality range—with the exception of the best and the worst sample. The training sequence gives the subjects the opportunity to get a feeling about the presented samples and the occurring disturbances.

The overall length of the listening test amounts to about 52 minutes, whereby two additional pauses of about 15 minutes are held in order to guarantee highly concentrated subjects over the whole test duration. After 20 to 25 minutes of continuous testing the subjects may get tired and the voting quality decreases.

Every single sample within the continuous stream starts with a short signal tone. Having re-played the voice sequence, a short pause of about five seconds is added in order to give time for the subject's decision process.

Test session	Sample length [s]	Number of			Overall length [s]
		Votes	Training samples	Total votes	
A	21	75	7	82	28 min 42 s
B	14	38	5	43	10 min 02 s
C1	13	38	5	43	9 min 19 s
C2	14	13	3	16	3 min 44 s
	Total	164	20	184	51 min 47 s

Table 4.4 – Subjects assess three different attributes of one conversation in four test sessions, which are distinguished by the corresponding sample length. The listening test lasts about 52 minutes.

4.3.4 Instructions for Subjects

Before starting the listening test, the subjects are asked, if they are right- or left-handed, if they suffer from hearing handicaps, if they use hearing aids, or if they are German mother-tongues or not. Normally, in the case of a wireline phone, right-handed people use the left hand to hold the telephone receiver to the left ear, since they use the right hand for other tasks such as taking notes. Moreover, people with hearing defects are excluded from the tests and, finally, only the votes of German mother-tongue subjects are taken into account, as the presented samples are also in German.

The basic introduction to the subjects starts with an overview on the time schedule including the different sessions and pauses. Then, the test setup is explained: The subjects are told that they are listening as a third party to a conversation between a male and female voice. They should put themselves in the place of the female speaker, who is conversing with the male user via a hand set. Thus, the subjects should assess how they perceive the male voice in a general quality sense and in terms of disturbances such as speech gaps. Moreover, the annoyance of their "own" female talker echoes is evaluated. Usually, in the case of a right-handed person, the male voice and the disturbances are only received in the left ear. When the sample is played back, the subjects only have a few seconds for voting.

Finally, the subjects are told that their votes can not be right or wrong, since the subjective impression is perceived individually by every single subject.

For every session a detailed explanation is given according to the occurring degradation. Since the third-party listening test focuses on the performance under double talk, the introduction, for example, for the C1 test session listed in Table 4.4 may look as follows:

> "The female talker is interrupted by the male user. How do you perceive echoes, especially when both participants are talking simultaneously?"

4.3.5 Voting Procedure

The voted attributes, which have already been mentioned in Figure 4.3, are either evaluated according to the ACR methodology or according to a modified version of the DCR principle (Table 4.5).

The first opportunity of sequence presentation is used to assess the overall quality of the transferred male voice samples. The second, modified technique uses a DCR voting scale according to Table 2.2 and presents the samples in an ACR test manner, because the pure DCR test according to [104] is only suited for little disturbances and leads to unreliable results, when the difference to the reference (of high quality) is high. The combined ACR and DCR methodology, which is used for the test sessions B, C1, and C2, is not standardized yet, but provides the commonly used concept for evaluating a broad range of impairments.

The subjects have been instructed to decide on one of the five opportunities of each opinion scale by placing a cross at the corresponding number in the questionnaire of the listening test. Furthermore, they also have been allowed to choose votes exactly in between two integer numbers (e.g., 1.5 or 3.5). The results of the third-party listening test are presented in Chapter 5.

Test session	Voted attribute	Methodology	Opinion scale
A	Overall quality of the male voice	ACR	Table 2.1
B	Drop-outs and speech gaps in the male voice	modified DCR	Table 2.2
C1	Echo-disturbance caused by the female voice		
C2			

Table 4.5 – Evaluated aspects of the conversational speech samples, the corresponding methodology of sequence presentation, and the listening scale for judgment.

Chapter 5

Results

"Statistics will prove anything, even the truth."[7]

5.1 Evaluation of Individual Votes

As already discussed in Section 2.1.6, the MOS values are obtained by averaging the single votes v_i of all N = 21 subjects:

$$\mathrm{MOS} = \frac{1}{N}\sum_{i=1}^{N} v_i. \tag{5.1}$$

Due to the limited number of subjects, and the 0.5 MOS granularity of the voting scale, which adds uncertainty representable as quantization noise, statistical properties such as confidence intervals have to be reported along with the MOS value [146].

Under the assumption that the MOS values according to (5.1) are drawn from a Gaussian probability distribution, and having chosen a significance level α (e.g., an α of 0.05 indicates a 95 percent confidence level), the confidence interval for the level α is derived from

$$\mathrm{MOS} \pm v(\alpha)\frac{\sigma_v}{\sqrt{N}}, \tag{5.2}$$

whereby σ_v stands for the well-known standard deviation of the v_i-population

---

[7] Noel Moynihan (1916–1994): British doctor and writer.

$$\sigma_v = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(v_i - \mathrm{MOS})^2}\,, \tag{5.3}$$

and v(α) denotes the range of the symmetric interval around the MOS value in which the area under the standard normal curve[8] equals (1 – α):

$$\mathrm{erf}\left[\frac{v(\alpha)}{\sqrt{2}}\right] = \frac{2}{\sqrt{\pi}}\int_0^{\frac{v(\alpha)}{\sqrt{2}}} e^{-t^2}dt = 1 - \alpha\,. \tag{5.4}$$

The computation of v(α) from Equ. (5.4) requires the inverse error function, which assumes a standard normal distribution; e.g., for $\alpha = 0.05$ this function results in $v(\alpha) = 1.96$. A higher confidence level (e.g., $\alpha = 0.01$ instead of $\alpha = 0.05$) yields larger confidence intervals. Furthermore, a higher variance in terms of the standard deviation of the individual opinion scores (i.e., less agreement among subjects) also results in larger intervals. Finally, increasing the number of listeners (without changing other parameters) decreases the symmetric intervals.

The Gaussian assumption is problematic for small sample sizes and, when the underlying distribution of individual scores diverges strongly from a normal distribution. In fact, the Gaussian function, used within this thesis, is sometimes replaced by the Student or by the t-distribution.

In this thesis, a 95 percent confidence level has been chosen. The single MOS results in the diagrams of the following sections are interpolated in order to demonstrate the general behavior more clearly.

5.2 Standard Double Talk Test

Echo cancellers are facing stringent requirements under double talk conditions, because the DTD always comes with detection errors. Consequently, the adaptive filter diverges and the NLP cuts off some syllables of the near end speech signals. The fixed parameter set of the DTD for the whole standard double talk test series corresponds to: $G_{DTD} = 0.707$ and $T_{hold} = 200$ ms (Table 4.4).

The diagrams in Figure 5.1 have been obtained for an unmodified male voice level ($A_v = 0$ dB), echo cancellers equipped with the NLMS filter, and the loud telephone—also see first line of "standard double talk" test in Table 4.4. The resulting graphs in Figure 5.1 a) represent a high level of overall male voice quality, since the near end signal has not been reduced as for the rest of the standard double talk test and, thus, the DTD works more reliable. The graphs of the high-quality results almost reach the average voting level of the completely echo-free reference sample. Moreover, both curves converge on each other from on about 50 ms. As the requirements on the overall echo attenuation of the new approach increase with higher echo delays and approach the high and constant TELR value of the conventional echo canceller in this way, the convergent behavior should be observed for every parameter combination. The standard approach delivers, in a mean sense, slightly

---

[8] The standard normal distribution comes with $\sigma = 1$ and a mean value of $\mu = 0$.

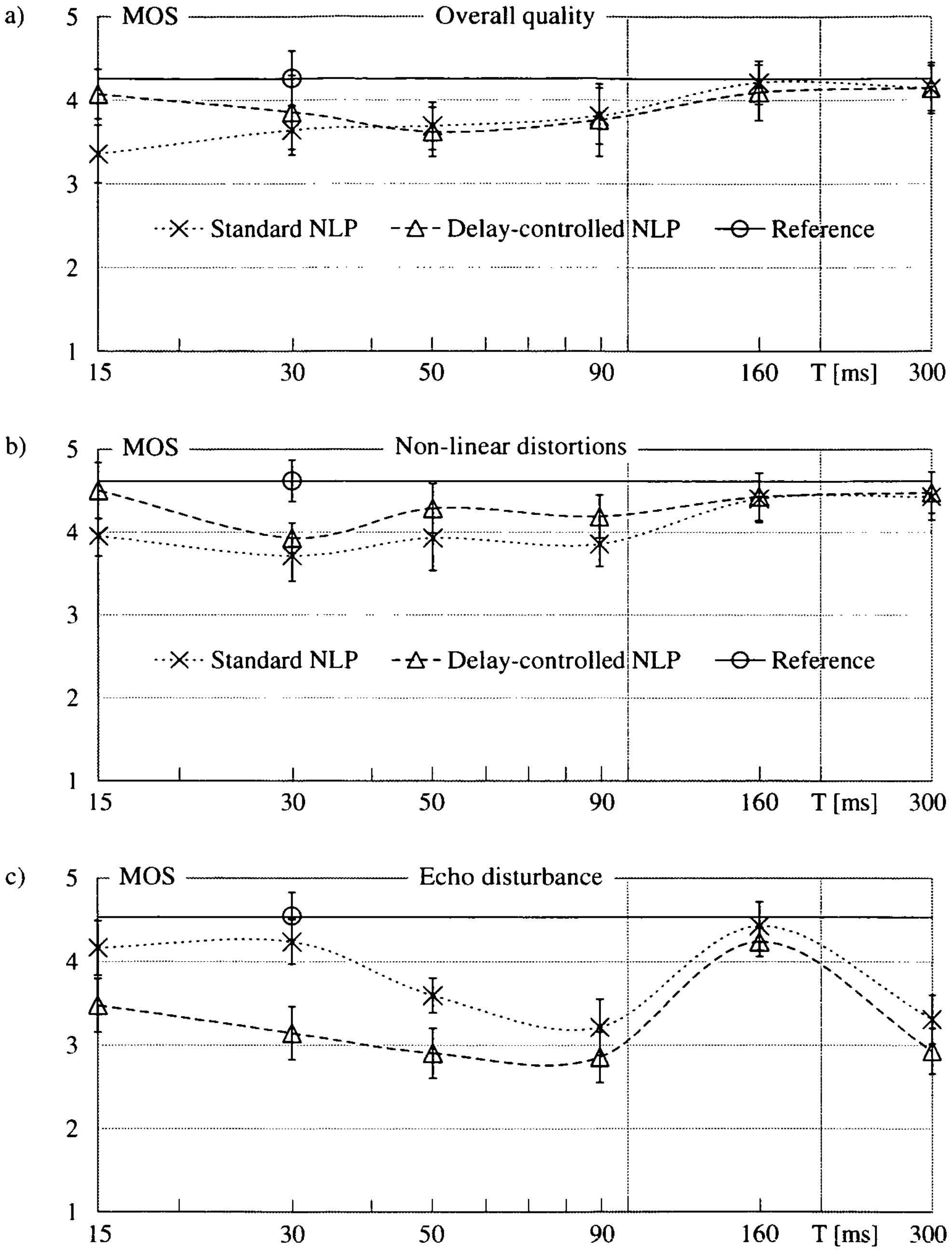

Figure 5.1 – Results of the standard double talk test using the adaptive filter, the original level of the male voice (A_v = 0 dB), and the loud telephone. a) The overall quality and b) the non-linear distortions present a high quality of the male voice. c) The disturbances caused by the female talker echo are more varying.

better results above 50 ms—but the confidence intervals of both graphs are almost totally overlapping. In this time range the delay-controlled concept admits hardly perceivable echo, which makes the performance very little worse. The impairments of the male voice introduced by the center clipping NLP occurs independently from the one-way delay—but the annoyance due to talker echo is getting worse with increasing delay. This behavior becomes more relevant for different settings of the experimental parameters, as illustrated in Figures 5.2 and 5.3. The improvement of the overall voice quality in terms of MOS difference between both graphs reaches its maximum at 15 ms and amounts to 0.7 MOS[9].

The results for the non-linear distortions, which are observed in the male voice, are illustrated in Figure 5.1 b). The curves point out better results in comparison to the evaluated parameter "overall quality" of the same sequences. The high average level of the votes originates from the fact that it is easier for test persons to concentrate on exclusively one aspect of the conversation such as speech gaps or drop-outs. Thus, a clear distinction between, for example, annoying and inaudible degradations is simpler. On the contrary, the attribute "overall quality" comes with additional effects such as echoes and clipping. The graphs of the standard and the delay-controlled approach are not continuously converging, but they finally coincide at 160 ms.

The outcomes for the assessed parameter "echo disturbance" under double talk conditions in Figure 5.1 c) show the following characteristics. Firstly, the standard approach performs, as expected, better than the new concept due to the allowed echoes in the latter case. But the gap of more than 1.0 MOS at 30 ms is unexpectedly high. Ideally, the achieved quality of the new concept should be at a quite constant level, because the control mechanism follows the talker echo tolerance curve. Moreover, echoes of full level are reflected back to the far end side in double talk periods detected by the DTD. The number and length of these special phases and the DTD detection behavior in such cases is hard to control and depends strongly on the speech material. A much broader range of samples from different communication sequences would have to be applied in order to obtain constant MOS values for the echo disturbance.

All in all, there has to be found always a trade-off between achievable near end voice quality and annoyance caused by female talker echoes. The higher the male voice quality, the more disturbing are the perceived echoes at the distant end and vice versa.

According to Table 4.4, the following Figure 5.2 is based on recordings made under worst case conditions, i.e., with a reduced level of the male voice (A_v = -6 dB) and without an adaptive filter (ERLE $\equiv$ 0 dB). The latter assumption stands for a divergent filter with reproducible behavior. In such systems only the NLP provides echo cancellation. Moreover, the handset with a reduced attenuation in the receiving direction, i.e., the loud telephone set, has been deployed.

The overall quality curves in Figure 5.2 a) converge from about 40 ms on to a very low level of about 1.5 MOS due to the bad environmental conditions. The degradations of the male voice, which are independent of the one-way delay, are dominating over the impairments due to talker echoes. Thus, the quality level of the conventional idea is kept quite constant over the whole delay range, while the delay-controlled concept shows this behavior only in the convergent state. The improvement expressed in difference of the MOS values achieves its maximum of 1.0 MOS at T = 15 ms. The reference samples, which have been recorded without any talker echo and without an echo canceller is always voted at a

---

[9] As often used in literature, the term "MOS" denotes both, the variable and the corresponding unit.

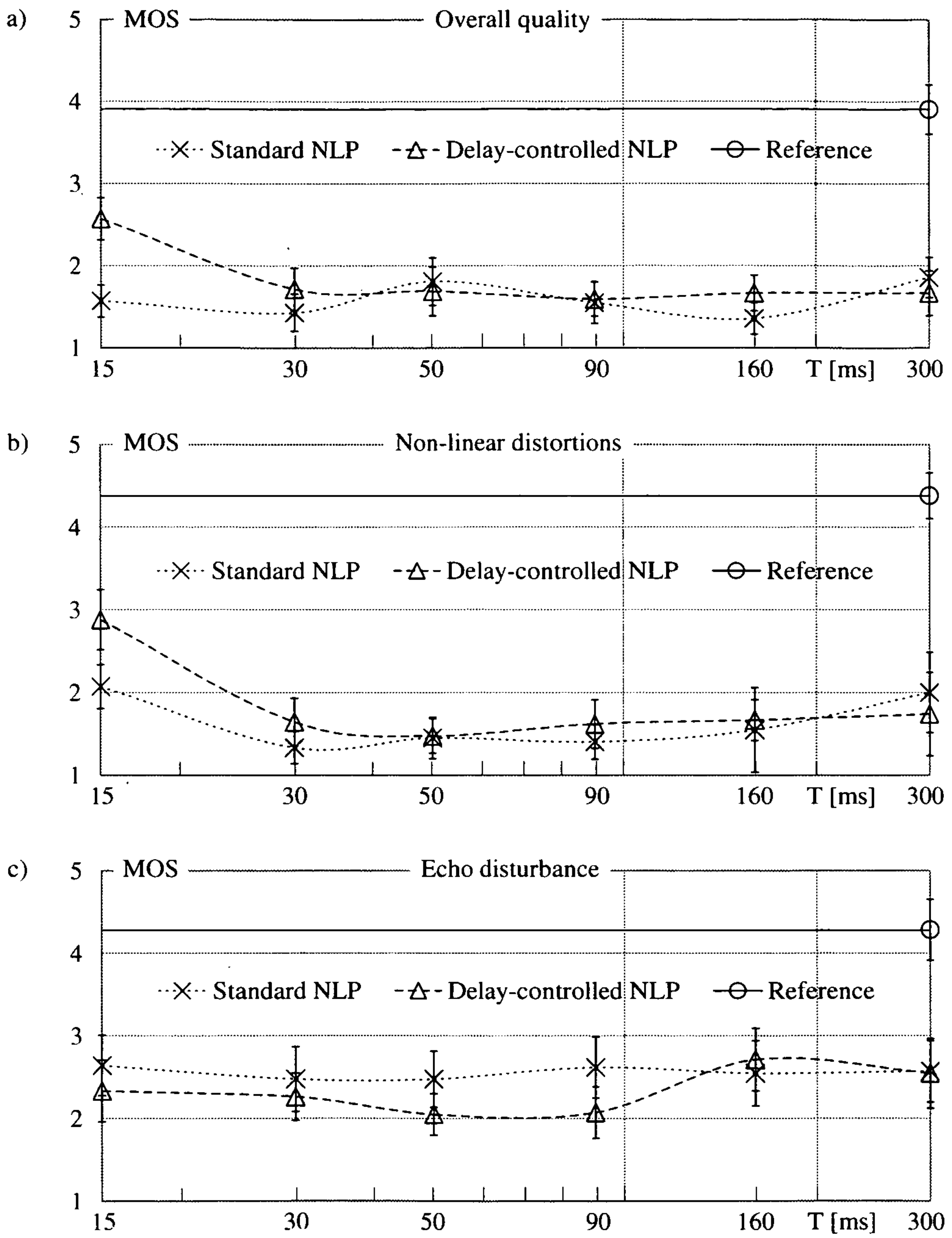

Figure 5.2 – Outcomes under standard double talk test conditions obtained without the adaptive filter, a reduced level of the near end signal (A_v = -6 dB), and by using the loud telephone receiver. a) Overall quality of and b) perception of speech gaps in the male voice. c) Echo disturbances caused by the female voice.

high level.

Since the annoyance of the non-linear distortions in the transferred male voice dominates the voting of the overall quality, the assessment of these special artifacts results in quite the same graphs, as depicted in Figure 5.2 b). The only difference is seen for

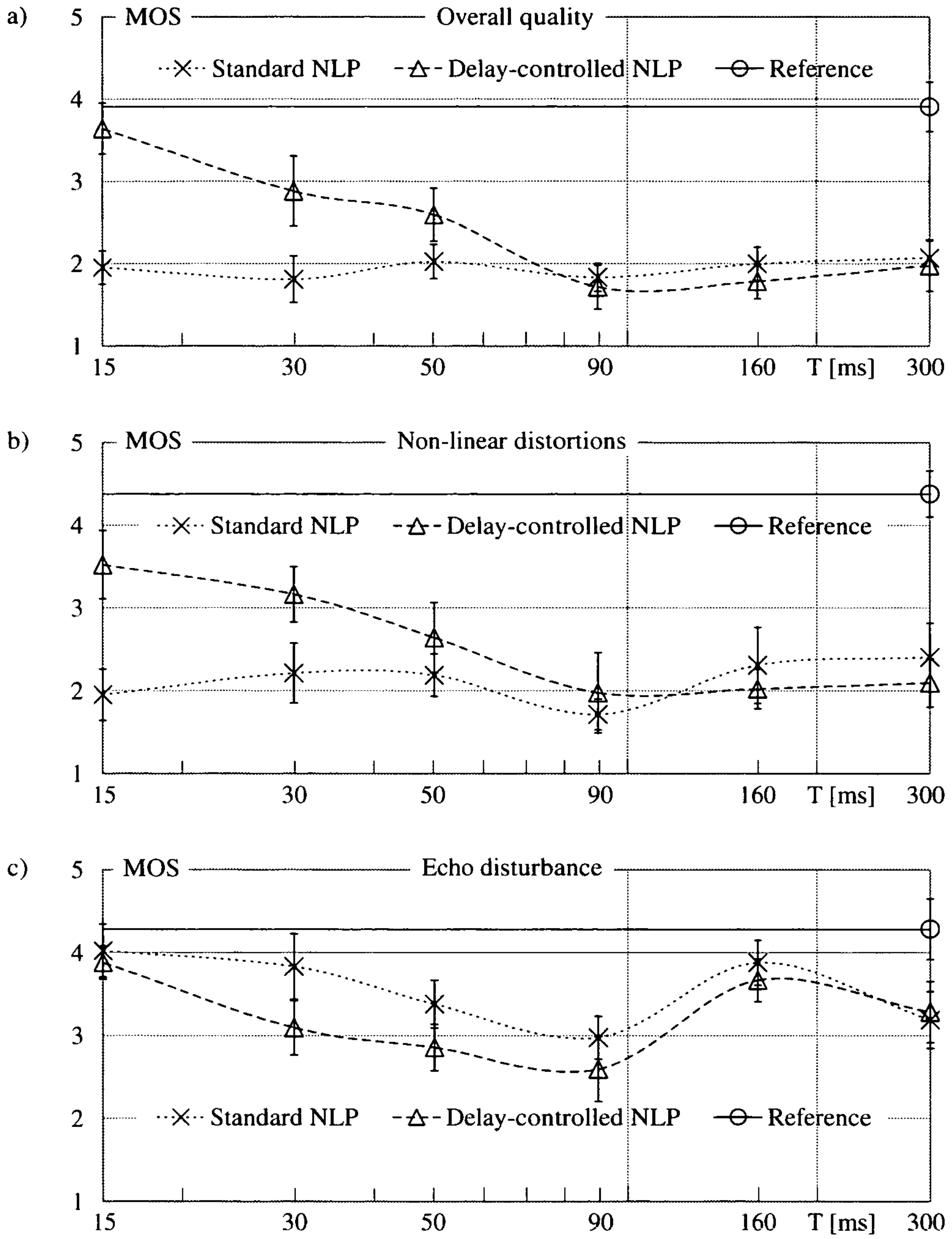

Figure 5.3 – Significant improvements of the conversational voice quality especially for low delay values in the case of an implemented filter algorithm, decreased male voice level, and the loud telephone set. a) Overall quality of and b) drop-outs in the male voice. c) Annoyance of the perceived female talker echo.

T = 15 ms, where the non-linear distortions are not that disturbing as the overall quality parameter indicates. This effect is especially observed for the traditional center clipping NLP.

Finally, the echo disturbance curves, which are illustrated in Figure 5.2 c), reflect more or less the expected functioning—i.e., quite constant quality levels and the standard application yields in slightly better outcomes than the new approach. The small deviation from this performance at T = 160 ms is of no statistical relevance.

The conditions of the third standard double talk test series, which are listed in Table 4.4, differ from the previous one only in the fact that the NLMS filter is deployed (Figure 5.3). Therefore, the improvements compared to Figure 5.2 are achieved due to the implemented filter algorithm [26] [27]. The results of the three reference levels in Figure 5.3 have been taken from Figure 5.2, since both have presumed A_v = -6 dB.

The new approach delivers better overall quality results, which are illustrated in Figure 5.3 a), especially for delay values up to 80 ms. The remarkable, maximum performance improvement of 1.7 MOS is reached at 15 ms. Although the male voice has been attenuated and the DTD performs quite badly, the delay-controlled concept almost reaches the undisturbed reference level of the echo-free connection in this case. Convergence is obtained at about 80 ms. The conventional NLP implementation performs quite constant around a MOS value of about 2.0 MOS.

The non-linear distortion graphs in Figure 5.3 b) are very similar to the corresponding one of the overall quality attribute, because the male voice is mainly degraded by the speech gaps and not by the talker echoes. The decrease of the conventional graph at 90 ms has not been expected and is a consequence of the deployed speech material.

The evaluation of the echo disturbance comes very close to the echo assessment in Figure 5.1, i.e., the new approach performs worse than the standard one. The quality gap between the two curves at 30 ms is a little smaller and reaches 0.7 MOS; it can be neglected for latency values of 100 ms and higher. As already mentioned before, these results are preliminary. A compromise between allowed echoes and degraded voice quality has to be found.

Finally, it is emphasized that Figure 5.3 points out very clearly the room for improvement of the delay-controlled NLP concept.

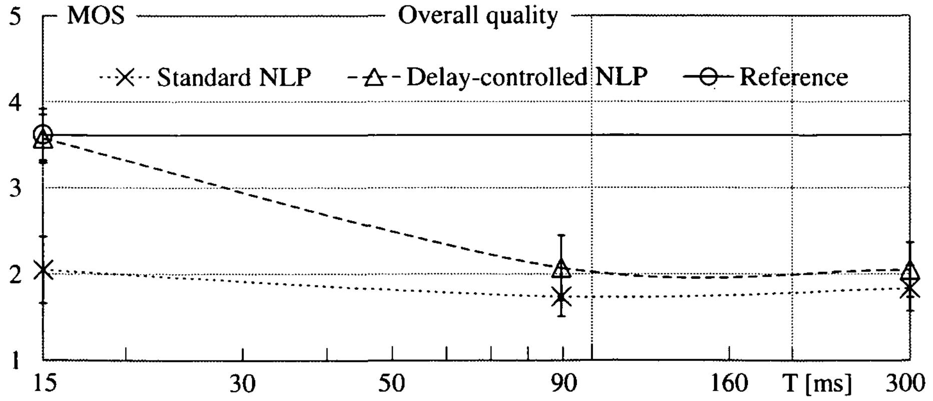

Figure 5.4 – Comparative investigation of the overall voice quality of the male participant deploying the standard silent ISDN phone "Europa 10" under equal settings as in Figure 5.3.

The loud telephone has been created with the intention to offer clearly audible speech samples to the subjects. In order to document the difference to the artificial loud telephone, the same set of parameters as applied for Figure 5.3 has been used for the original, silent phone. The outcomes of the assessed attribute "overall voice quality" are shown in Figure 5.4. The graphs of both approaches in this diagram are almost identical to that one of the loud telephone; the maximum difference of the delay-controlled approach between both types of telephones amounts to 0.36 MOS at a one-way delay of 90 ms, while the conventional approach shows a maximum variance of 0.24 MOS at 300 ms. The curves in Figure 5.4 show a smoother behavior than that one of Figure 5.3. The modification of the standard ISDN phone has made the distortions more audible, but it has not significantly changed the perceived voice quality. Again, substantial improvements of up to 1.5 MOS are achieved when deploying the delay-controlled echo canceller.

All the presented results are preliminary and need to be further analyzed. The new approach could certainly be tuned leading to echo attenuations comparable to standard implementations without loosing the advantage of minimized double talk impairments.

5.3 DTD Optimization

This optimization test series has been conducted with the intention to derive optimal parameter settings for the DTD in terms of the relative level threshold G_{DTD} and the holding time T_{hold}. The following general framework points out the influence of the DTD very clearly, since the detection errors are increasing with decreasing near end voice levels (A_v = -6 dB). Moreover, the disabled filter allows high talker echo components, which are perceived at the far end. Finally, the loud telephone set with RLR = 0.68 dB has been utilized for the recordings.

Beside the resulting graphs of the five different parameter sets, which are listed in Table 4.4, the curves of the standard DTD values (G_{DTD} = 0.707, and T_{hold} = 200 ms) are drawn to show the differences in terms of the assessed parameter "overall quality". The standard curves are also found in Figure 5.2—they only differ in the reduced number of used one-way delay values. Furthermore, all five diagrams contain the average vote of the same reference sample, because they are all based on a reduced near end level and a loud telephone set is utilized. The confidence intervals are omitted in the diagrams in order to provide clearly arranged graphs.

On the one side, Figure 5.5 points out the behavior of the traditional concept and the delay-controlled approach for varying DTD holding times and a constant G_{DTD} parameter of 0.707. On the other side, T_{hold} is kept at a constant level of 200 ms in Figure 5.6, while the relative attenuation factor G_{DTD} is altered.

Figure 5.5 a) represents the achievable voice quality for a reduced DTD holding time of T_{hold} = 30 ms. Both NLP types point out a worsened performance compared to the standard parameter settings for all three one-way delay values due to the fact that the level-based Geigel algorithm, which has been deployed for the creation of the voice samples, decides on double talk only in distinct and very short periods of time. Especially for low near end signal levels the degradations due to filter divergence and NLP clipping are increasing

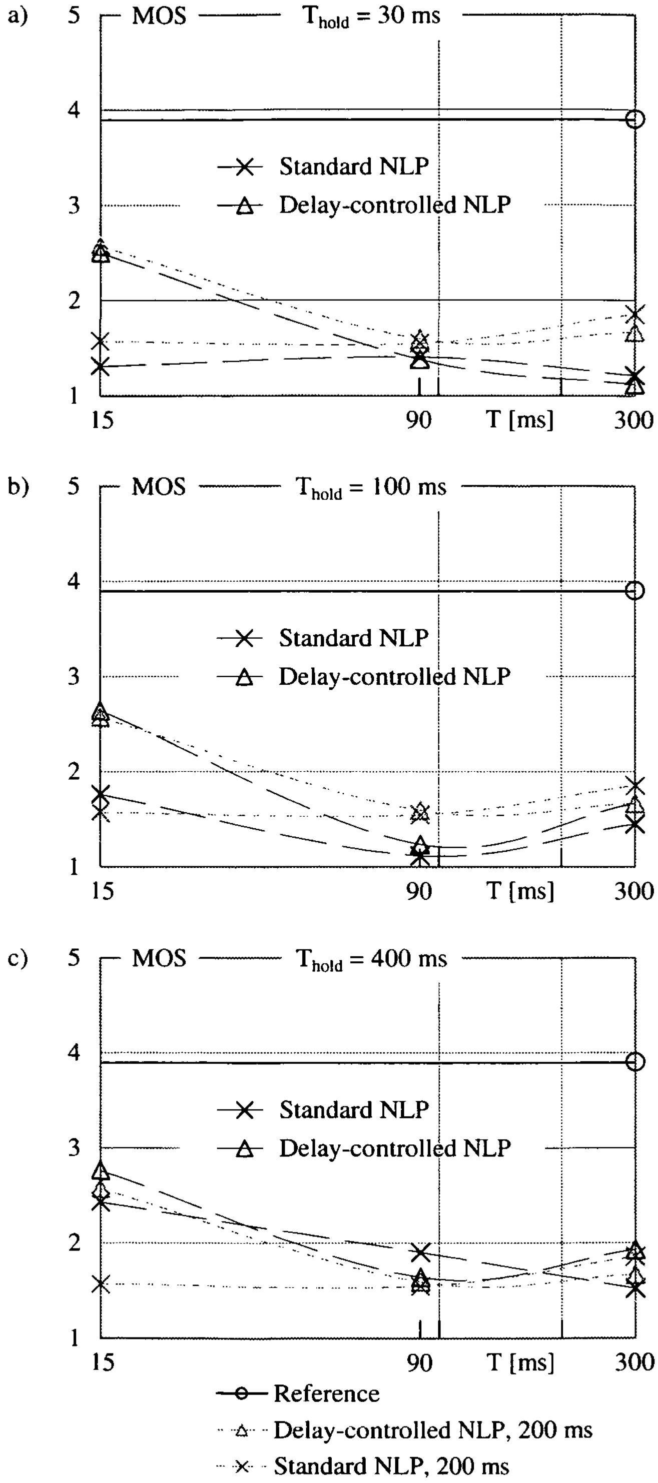

Figure 5.5 – Influence of varying DTD holding times T_{hold} on the overall voice quality under critical parameter combinations: Divergent filter condition (assuming no filter with ERLE ≡ 0 dB), and low voice level induced at the male participant's phone (A_v = -6 dB). The loud telephone is deployed for the recordings. a) T_{hold} = 30 ms. b) T_{hold} = 100 ms. c) T_{hold} = 400 ms.

when decreasing the holding time, as more and more undetected double talk periods are occurring.

When implementing the DTD with T_{hold} = 100 ms, the perceived overall voice quality in Figure 5.5 b) shows an impaired performance for 90 ms. The delay-controlled approach delivers almost the same results for the other two delay values. On the contrary, the standard NLP approach has a much more varying quality level over the one-way delay. At 15 ms the performance is slightly improved, and for T = 300 ms the voice samples are judged with lower quality.

Finally, a quite high DTD time of T_{hold} = 400 ms leads to an improved behavior in terms of the attribute "overall voice quality". The delay-controlled approach performs slightly better over the whole time range than the reference setting, which comes with T_{hold} = 200 ms. The standard concept improves the voice quality only for 90 ms. The advanced performance stems from the fact that with enlarged DTD holding times more wanted voice samples are transferred unchanged over the network to the far end. Generally, the echo canceller is seen as completely transparent for the voice signals under the chosen circumstances in that case. Thus, the undamped echo periods are getting longer with increasing holding times, since they are reflected unmodified via the NLP, when the DTD disables the non-linear unit in double talk periods. As the clipping effects and, thus, the disturbing drop-outs in the male voice are stronger than echoes under the assumed conditions, the voice quality improves slightly. Under more relaxed conditions the occurring echoes would be more annoying and the overall quality level decreases.

Recapitulating it is stated that T_{hold} = 200 ms provides the best choice of all four T_{hold} variations for the DTD.

Compared to the standard setting with G_{DTD} = 0.707, Figure 5.6 a) is based on a relative DTD level parameter of G_{DTD} = 0.625. The resulting graphs are almost identical for both perception oriented systems, while the standard parameter choice results in better results for T = 15 ms for the traditional NLP. Therefore, the DTD parameter may be chosen according to the investigated value of G_{DTD} = 0.625, which makes the detection behavior of the DTD a little less sensitive for double talk phases.

The male voice quality is degrading with an increasing relative level parameter. Figure 5.6 b) presents the corresponding results for G_{DTD} = 1. The higher the chosen G_{DTD} value, the more double talk periods remain undetected, and, consequently, the number of non-linear artifacts in the male voice is getting higher, while echoes are cancelled more sufficiently. The conventional approach for G_{DTD} = 1 performs better than for the standard setting for a one-way delay value of 15 ms, because large periods of the male voice are completely suppressed during the long double talk phase. The subjects have preferred totally missing male words in comparison to temporarily clipped speech segments and, thus, provided higher votes, although essential parts of the information have been missing. A real-world implementation would, in this way, not satisfy the required needs, because of the latter fact.

All in all, the optimum relative parameter is found around G_{DTD} = 0.625.

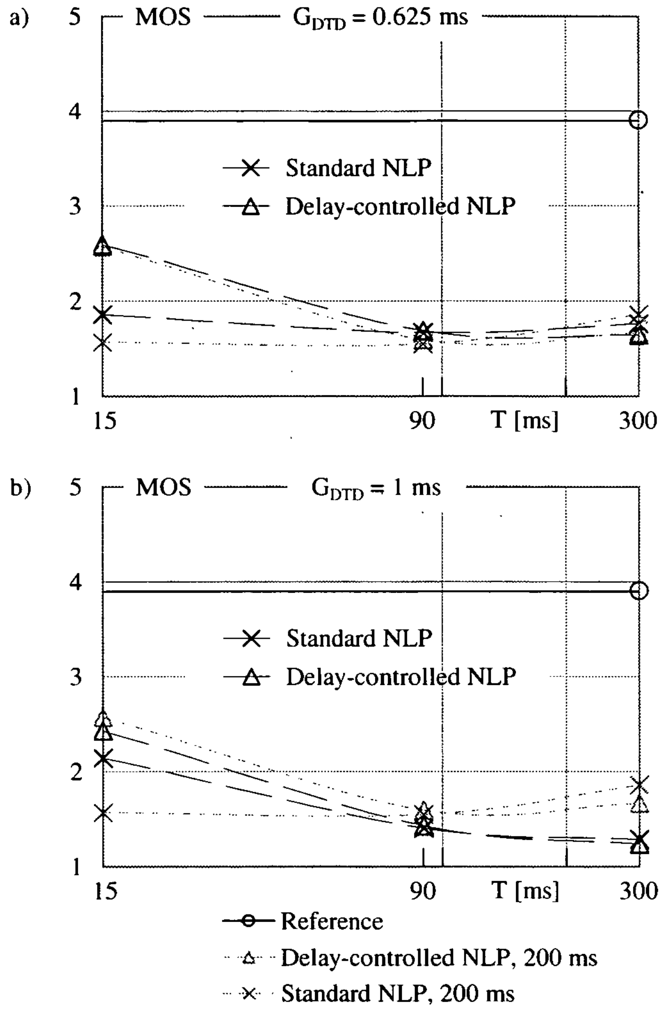

Figure 5.6 – Variation of the relative level parameter G_{DTD} and the consequences on the male voice quality perceived at the far end. The echo canceller comes without the adaptive filter and with a decreased male voice level (A_v = -6 dB). a) G_{DTD} = 0.625. b) G_{DTD} = 1.

5.4 Echo References

The purpose of this test series is to provide an anchor for the other results of the third-party listening test. In fact, the near end signal is completely suppressed, and the echo canceller is removed from the network (also see Figure 4.8). Consequently, the talker echo references are evaluated under far end single talk conditions. The settings of the echo references test correspond to the lower part of Table 4.5. The hybrid attenuation ERL is made variable in order to meet the required TELR values along the talker echo path. On the one hand, the "acceptable" and the "limiting case" graph of the talker echo tolerance curve of Figure 2.16 are examined. On the other hand, variable TELR values are taken according to Figure 4.9 in order to provide a vertical scan at constant one-way delay values. The speech samples have been created for end-to-end delay values of 15 ms and 300 ms. The assessed voice sequences have been recorded by using the artificial, loud telephone. Finally, the attribute "disturbances introduced by female talker echo" has been evaluated for both test series.

Ideally, the resulting graphs of the talker echo tolerance curves in Figure 5.7 should be horizontal lines; but there is a small descent of about 1.0 MOS for the "acceptable" and of around 0.8 MOS for the "limiting case" curve over the whole time range [27]. The decreasing score values for high echo delays are originating from two reasons: Firstly, they are depending on the utilized voice sequence, since the masking of important echo components varies with them. Secondly, the reduction of voice quality from about 200 ms on has already been observed in another listening only test [119]. Therefore, the talker echo tolerance curves are likely not representing the promised, constant level of echo disturbances. Further investigations based on a broad range of different voice samples would have to be carried out in order to correct the corresponding curves. Having found more reliable curves, the influence of the special voice sequence utilized for the third-party listening test of this thesis is quantifiable. The displacement between the curves in Figure 5.7 remains constant at about 0.8 MOS, which matches the expectations. Moreover, a high quality is observed for the "acceptable" graph. In addition to the standard assumption for the loudness ratings of SLR + RLR = 10 dB, the control unit has also used the exact values of SLR + RLR = 8.68 dB according to Table 4.1. These further assessments have been undertaken to show the introduced error in terms of voice quality, when the NLP control algorithm makes an assumption on the sum of loudness ratings of the deployed telephone. In closed communications networks, the system engineer may have knowledge on the phone's loudness parameters, while these values are not available in other arrangements.

In addition to the parameters listed for the "variable TELR" test in Table 4.5, the results obtained for the "talker echo tolerance curves" test have also been used for the graphs in Figure 5.8. The abscissa represents the true TELR values experienced along the talker echo path. This value is derived from the TELR objective of Table 4.5 by subtracting the

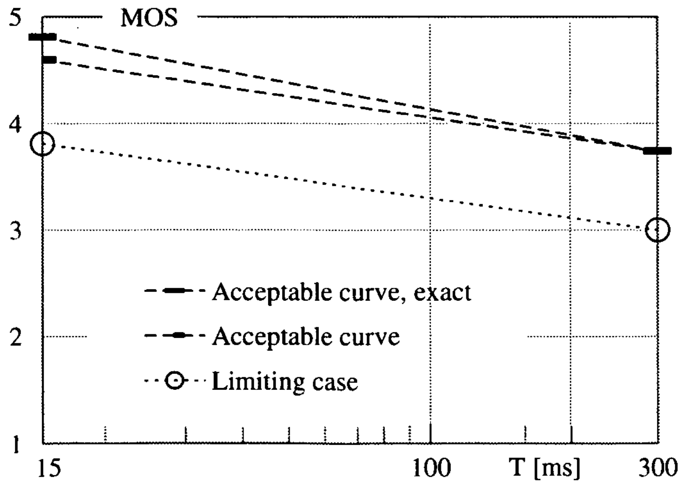

Figure 5.7 – Scan of the standardized talker echo tolerance curves in order to give reference MOS values for the whole listening test. The "acceptable" and the "limiting case" curve are evaluated under far end single talk conditions. Moreover, the control mechanism is provided with the exact loudness ratings of the deployed artificial telephone in the "acceptable" case.

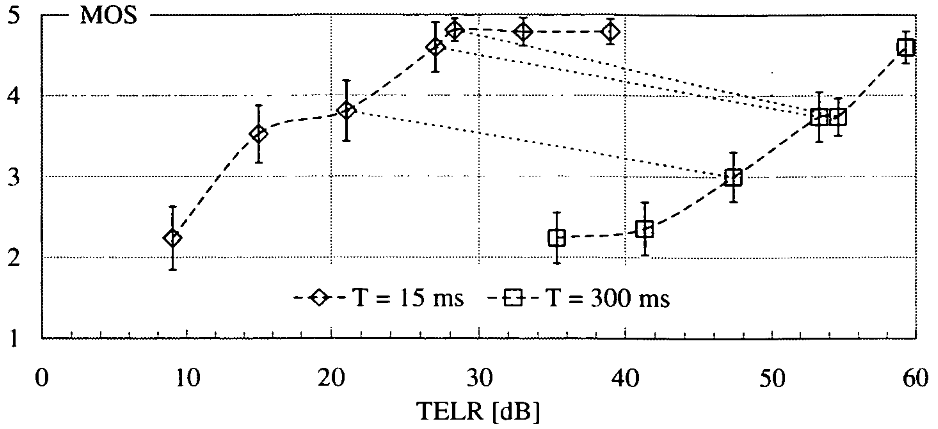

Figure 5.8 – Vertical scan of the talker echo tolerance curves at constant one-way delays and illustrated over the real TELR value along the echo path. The three connecting lines indicate the corresponding talker echo tolerance curves enhanced by the exact loudness ratings used for the NLP control mechanism.

difference between the assumed sum of loudness ratings and the present one of the deployed phone (i.e., the difference equals 10 dB – 8.68 dB = 1.32 dB for the loud telephone set according to Table 4.1). This deviation corresponds to the control error in terms of attenuation in decibel introduced by the NLP control algorithm. The graph of the 15 ms curves saturates at TELR values of about 28 dB, i.e., the perceived quality cannot be improved over the very high value of 4.8 MOS. The attenuation for the low one-way delay of 15 ms is high enough in this range so that the echoes are not disturbing any more. Moreover, the uncertainty of the votes in terms of the confidence interval is getting narrower. The slopes of both graphs—the 15 ms and the 300 ms—in the not saturated area are almost identical. Additionally, the outcomes for the two corresponding TELR values of each curve—the "acceptable", the "limiting case", and the exact one—are connected in pairs. Those graphs show again a descending tendency over the TELR. As already seen in Figure 5.7, the MOS values corresponding to the exact sum of loudness ratings (i.e., 8,68 dB) are evaluated with a higher quality than the standard ones (which are based on SLR + RLR = 10 dB), since the reduced sum of 8,68 dB requires higher values for the NLP attenuation in order to meet the requirements of the "acceptable" curve.

5.5 Conclusion

The outcomes of the third-party listening test, which compare the standard with the newly developed delay-controlled echo canceller, are summarized as follows:

The parameter "overall voice quality" points out improvements under the critical double talk test conditions especially for low delay values. Moreover, the graphs of both types of NLP implementations converge at mean end-to-end delay values of about 50 ms. A remarkable increase in terms of perceived voice quality is documented in Figure 5.3; it is achieved for a reduced near end voice level and for an echo canceller with an implemented adaptive filter. In this special case, the graphs, which correspond to both types of implementations, converge for delay values above 90 ms.

Since the non-linear distortions, which are caused by the NLP due to detection errors of the DTD, are dominating in the voice samples, the attribute "non-linear distortion" follows more or less the graph of the corresponding overall quality curves.

The evaluation of the attribute "echo disturbance" shows some deviations from the expectations, since there two types of echo signals have been occurring in the received voice stream. Firstly, the adaptive filter has to some extent attenuated the hybrid reflections (which are very high due to the worst-case assumption of ERL = 6 dB) and provided, quite low echoes to the far end talker. Secondly, in the case of double talk, the adaptive filter has diverged due to detection delays, and the NLP is disabled in such talking periods. The latter fact is responsible for undamped echoes, which are perceived very disturbingly. The amount and duration of those periods strongly depends on the speech material and parameter settings. Moreover, the echo disturbance varies with the one-way delay, because the masking of the echoes by the near end signal is differing and, moreover, the "acceptable curve" almost certainly provides TELR values, which are responsible for decreasing MOS values. The latter fact has been found out in the corresponding echo references test, which is presented in Figure 5.7. All in all, a compromise between occurring echo components and improvements of the male voice quality has to be found.

The error introduced by the NLP control's assumption on the loudness rating is kept low, as shown in Figure 5.7.

The DTD optimization test series pointed out that the corresponding parameters have been chosen well. Further improvements are achievable when applying a broader range of speech material and conducting another listening test.

Chapter 6

Outlook and Summary

Although available DSP computation power is always increasing and available at a reasonable price, echo cancellation will always be a topic of interest as long as there are two-wire customer loops connected to the backbone via hybrids. Especially in the case of an ever emerging IP network, the basic principle of coding and packetization always comes with a substantial amount of delay. The overall extent of the experienced time span from the talker's mouth to the talker's ear is reduced and, hence, the echo problem is diminished, when voice traffic prioritization schemes such as MPLS are applied.

Based on the theoretical concept and on the conclusions of the third-party listening test presented within this thesis, the next step would be to implement the delay-controlled enhancement upon a commercially available echo canceller.

Firstly, decisions have to be made as to which delay measurement method is suitable for such a prototype. Since the newly introduced idea is well suited for closed communication networks with one-way delay values below or around 100 ms (also see Figure 5.3), a slightly modified and enhanced version of the RTP, which acts similarly to the NTP, would satisfy the requirements for the determination of the round-trip delay values. Alternatively, a satellite receiver system would guarantee synchronized clocks with high precision at both media gateways involved. Thus, one-way delay values could be evaluated for every transferred IP packet with sufficiently low uncertainty. In addition to the GPS, proprietary packets for exchanging the gathered time information between the delay measurement units of the two echo cancellers would be required. Having implemented a delay measurement system, the verification of such an entity in terms of deviation of the real mouth-to-ear delay should be verified by injecting test traffic at the telephone receiver and measuring the delay of the reflected talker echo by filtering out the echo signals. Such a measurement would have to be carried out under different IP network conditions, since the experienced time span by the IP packets may vary considerably with the current load on the network.

Based on the resulting error of the deployed measurement unit, the uncertainty in terms of deviation from the required TELR value is determined according to Figure 3.4. Besides the time span experienced along the network, the delays introduced by the gateway itself and the succeeding PSTN have to be added in order to correct the results.

Secondly, the continuous determination of the attenuation of the received echo signals (i.e., the value of ERL + ERLE) has turned out to be another challenging issue. The accuracy of the measurement is decreased by an unreliable detection of double talk periods, because in such special talking situations a short detection delay of the DTD causes the control unit to regard the corresponding amount of near end signal as echo components and, in this way, falsifies the calculation of echo attenuation along the hybrid and the adaptive filter (also see Figure 4.5). Moreover, the adaptive filter diverges, and some echoes arrive unaltered at the far end talker via the disabled NLP in the case of undetected double talk. All in all, a high detection reliability of the DTD improves the overall results considerably. The methodologies suggested within this thesis either quantify the total sum of echo attenuation in one step according to Equ. (4.22) or both addends are calculated separately as described by Equs. (4.21) and (4.23). The latter concept tends to provide better results—as shown by the corresponding signals of the simulation results in Figure 4.5; but this has still to be confirmed in real-world scenarios.

As a next step the talker echo tolerance curve, which the new approach on echo cancellation relies on, is fine-tuned. The system engineers have to find a compromise between the improvement of the near end voice quality and how much echo is reasonable for the talker.

Finally, the only way on how to deduce the improvement of the system behavior due to the delay-controlled enhancement is to evaluate the subjective impression of the users by conducting another third-party listening test based on the prototype. Alternatively, the deployment and use of the new system in the real-world would carry out the overall performance satisfactorily.

Appendix A

Simulation Models

A.1 Overall System

In a first step, the models of the overall network echo canceller deployed in a mixed analog and IP environment are presented. Every diagram includes both, the standard as well as the delay-controlled simulation model. The composition of the diverse functional entities, which are colored gray, is explained in Section A.2.

A.1.1 With Adaptive Filter

The diagrams of Figure 4.5 have been obtained for the real-world implementation of the NLP control unit (Figure A.1), which correspond to Equs. (4.21) and (4.23), as well as for an ideal measurement based on Equ. (4.18). The former model is depicted in Figure A.1, while the latter one is illustrated in Figure A.2.

A.1.2 Without Adaptive Filter

The corresponding model of Figure A.3 is very similar to the ideal one of Figure A.1; the NLP control unit is even more simplified.

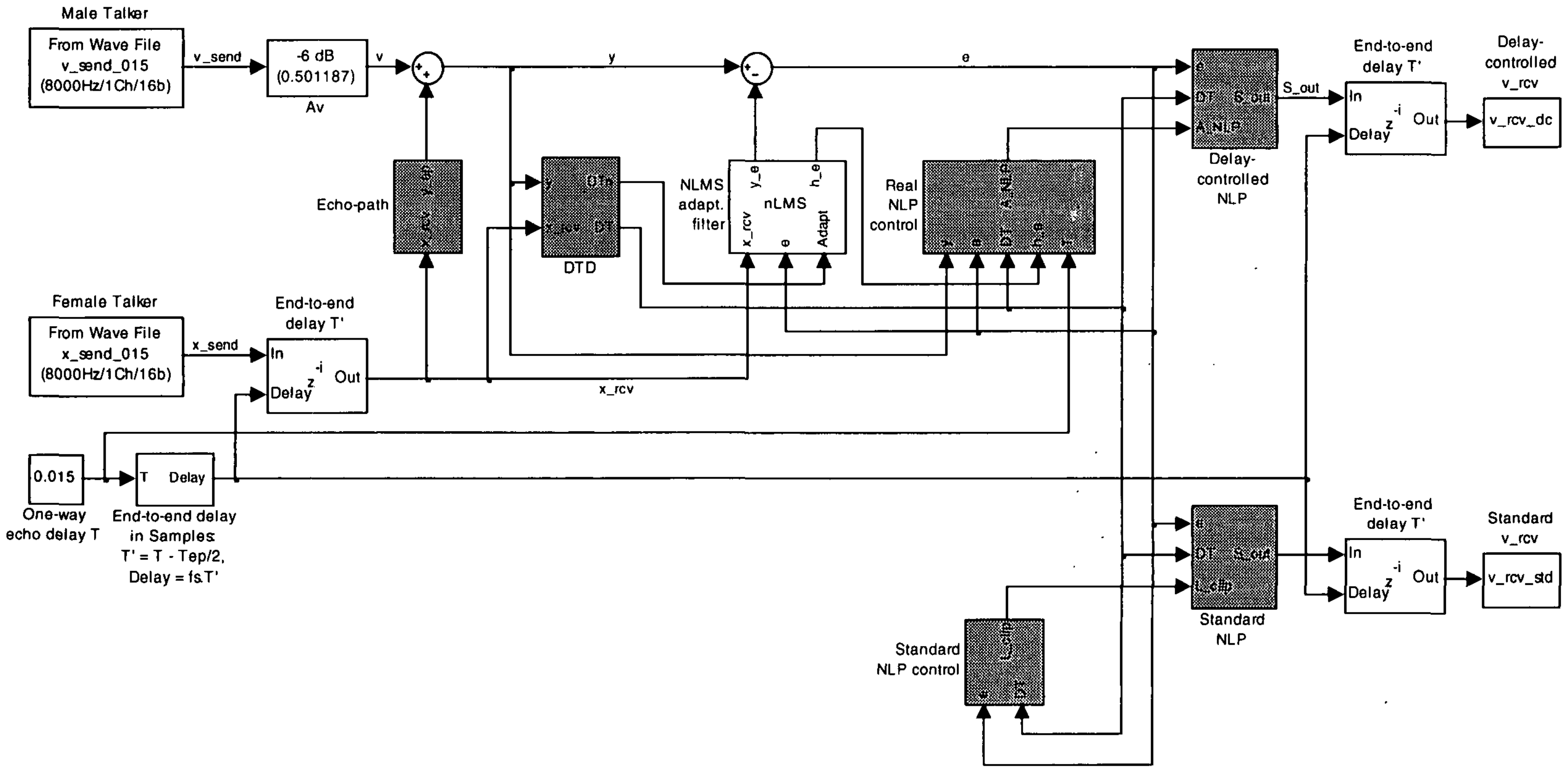

Figure A.1 – Simulation model of the real-world NLP implementation.

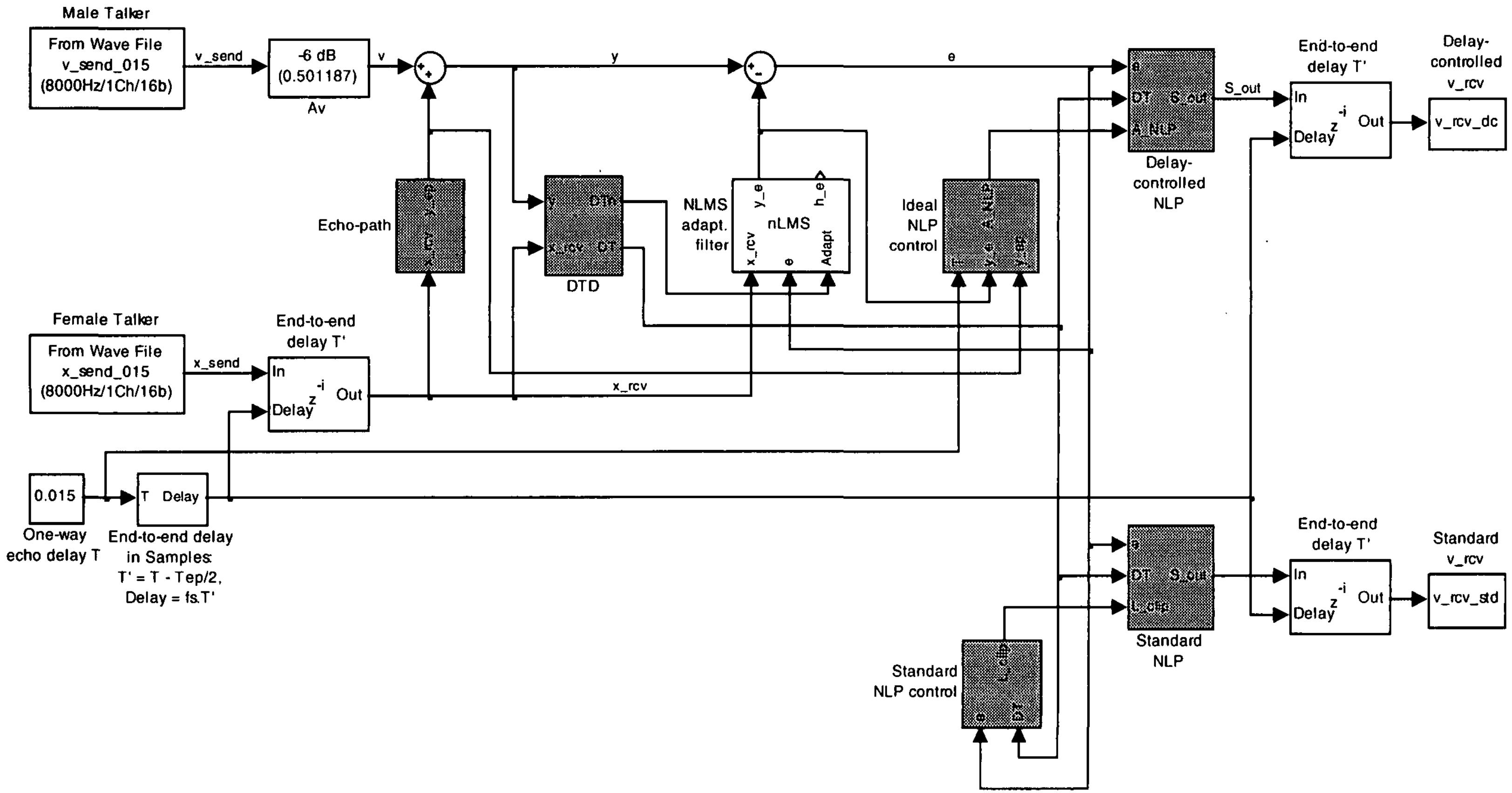

Figure A.2 – Simulation model of the ideal NLP implementation.

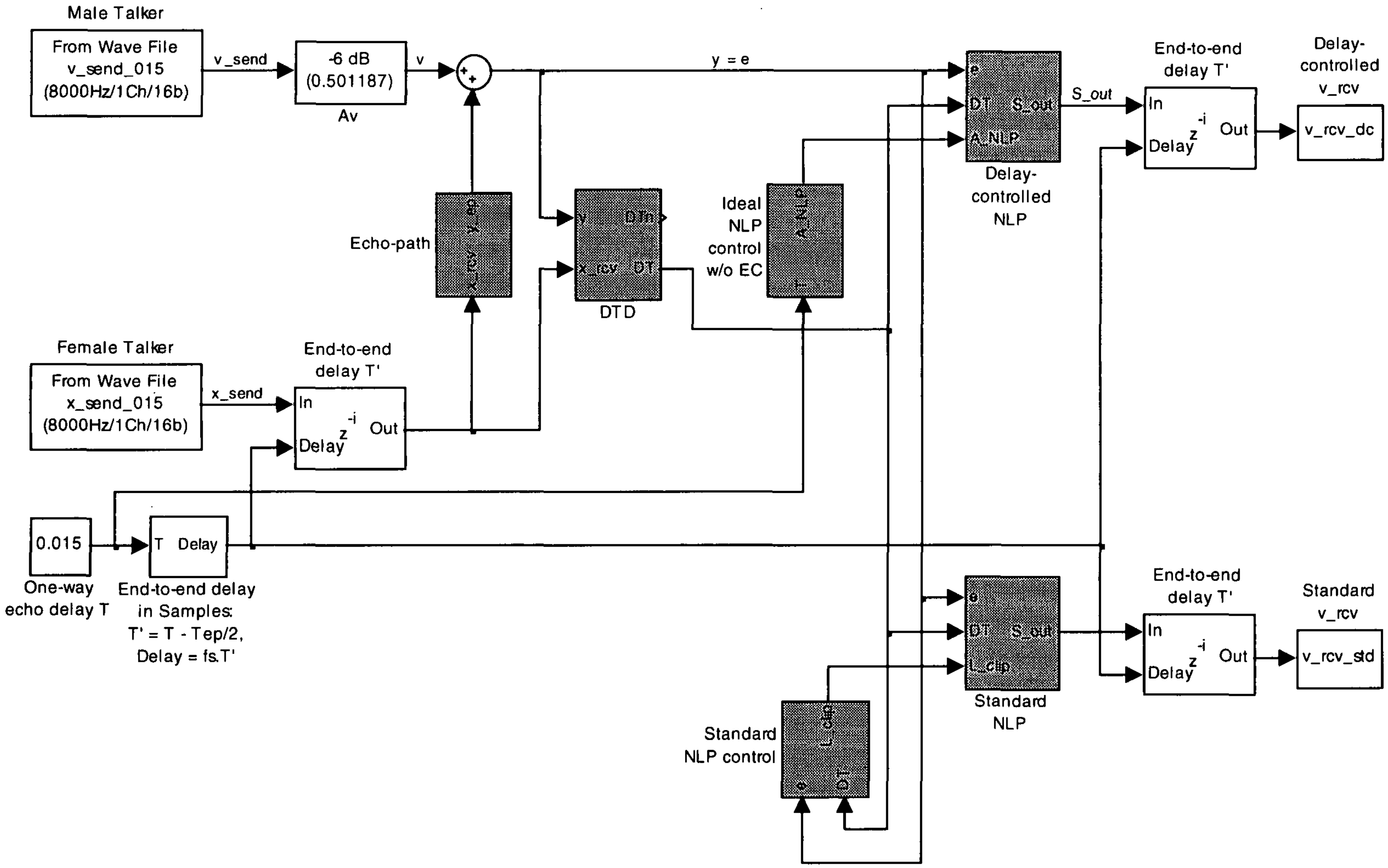

Figure A.3 – Ideal simulation model without the adaptive filter.

A.2 Echo Canceller and Echo-Path

The model of the NLMS algorithm has been taken from the Matlab® DSP Blockset library, while the other models have been designed according to the requirements.

A.2.1 DTD

The DTD has been realized according to Equ. (4.6). The corresponding block diagram is given by Figure A.4.

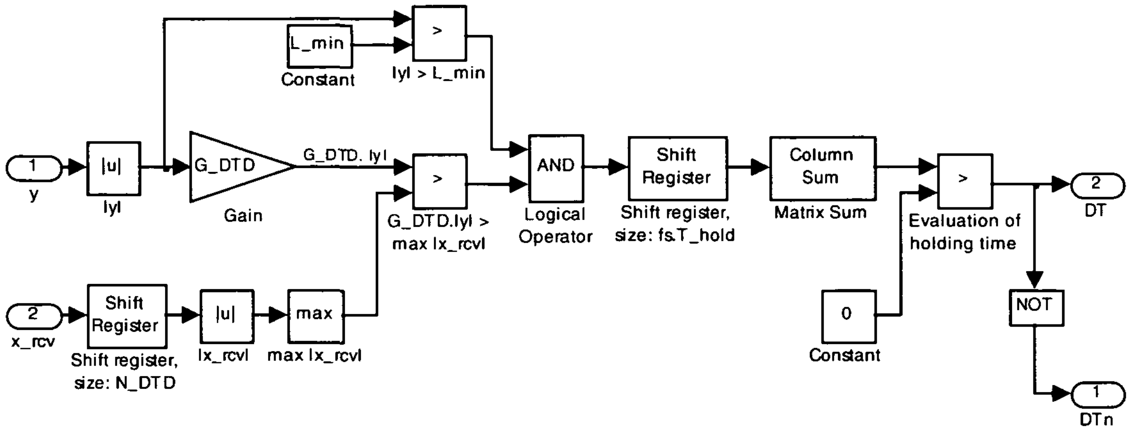

Figure A.4 – Simulation model of the DTD algorithm.

A.2.2 NLP and NLP Control Mechanism

The delay-controlled as well as the standard implementation of the NLP, which are presented in Figure A.5 and Figure A.6, respectively, are realized identically for the three overall systems of Section A.1.

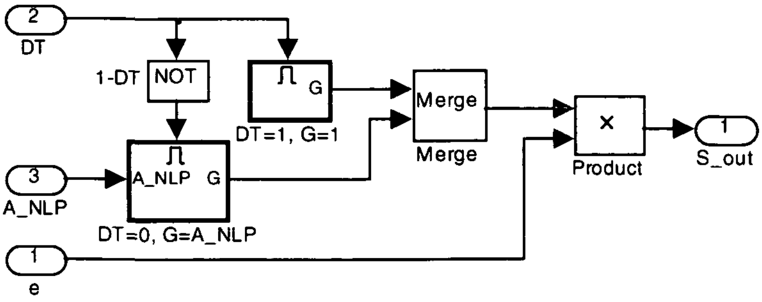

Figure A.5 – Simulation model of the delay-controlled NLP.

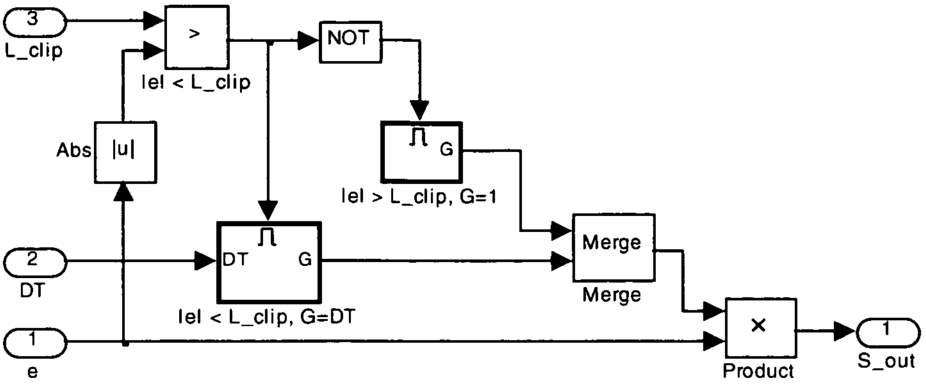

Figure A.6 – Simulation model of the standard NLP.

The block diagram of the real NLP control unit of Figure A.1 is made up of the following functional components, which are depicted in Figure A.7. The simulation functions of the "attenuation measurement" and of the "linear NLP attenuation" block are shown in Figure A.8 and Figure A.9, respectively.

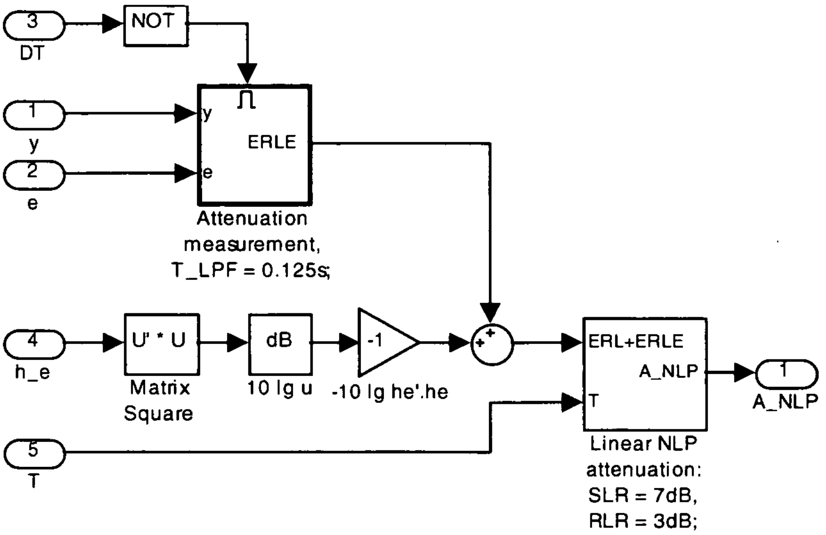

Figure A.7 – Simulation model of the real NLP control unit.

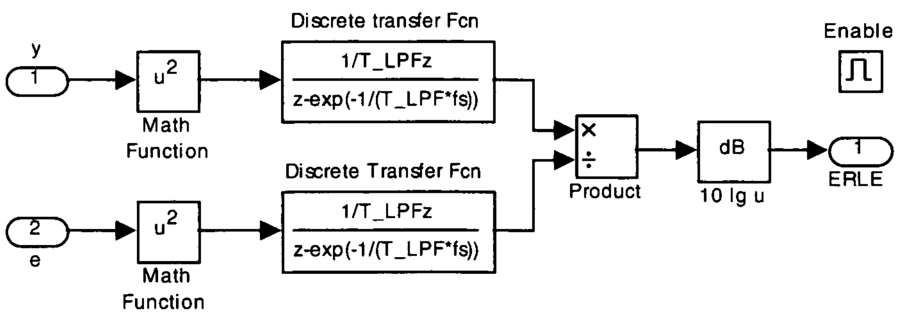

Figure A.8 – Simulation model of the "attenuation measurement" block in Figure A.7.

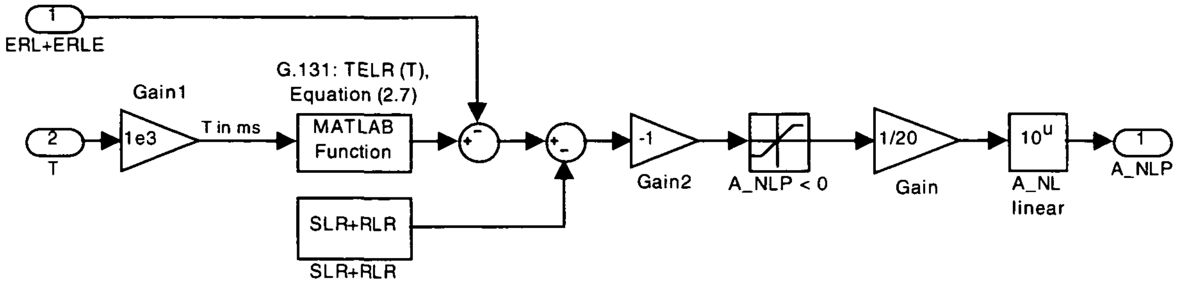

Figure A.9 – Simulation model of the "linear NLP attenuation" unit in Figure A.7.

The set-up of the conventional NLP control unit is described by Figure A.10 and Figure A.11, while Figure A.12 shows the ideal NLP control unit.

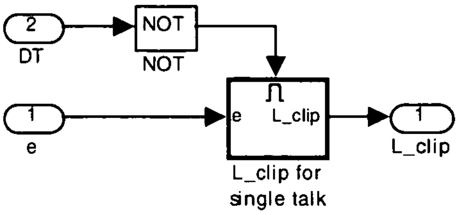

Figure A.10 – Simulation model of the conventional NLP control entity.

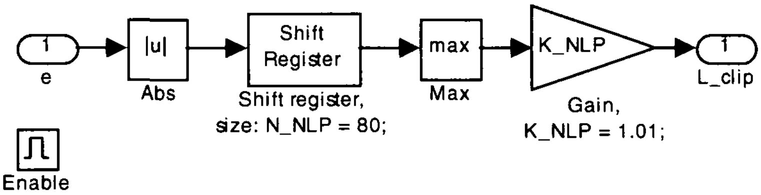

Figure A.11 – Simulation model of the "L_clip for single talk" entity in Figure A.10.

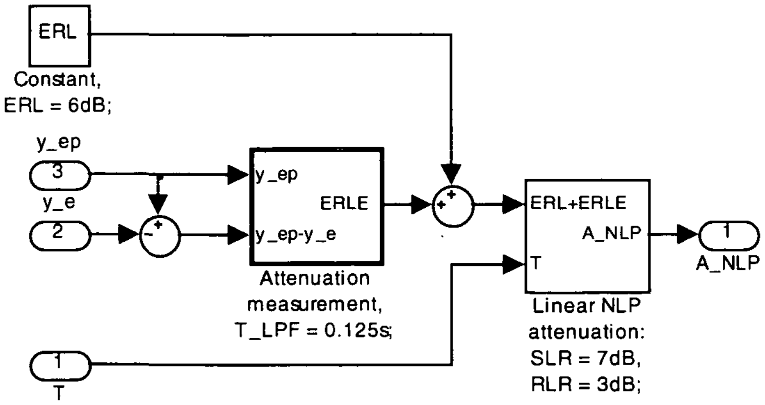
ERL
Constant, ERL = 6dB;
y_ep
3
y_e
2
y_ep
ERLE
y_ep-y_e
Attenuation measurement, T_LPF = 0.125s;
ERL+ERLE
A_NLP
T
Linear NLP attenuation: SLR = 7dB, RLR = 3dB;
1
A_NLP
1
T

Figure A.11 – Simulation model of the ideal NLP control unit in Figure A.2.

The NLP control mechanism for the echo canceller with a disabled adaptive filter according to Figure A.3 is shown in Figure A.12.

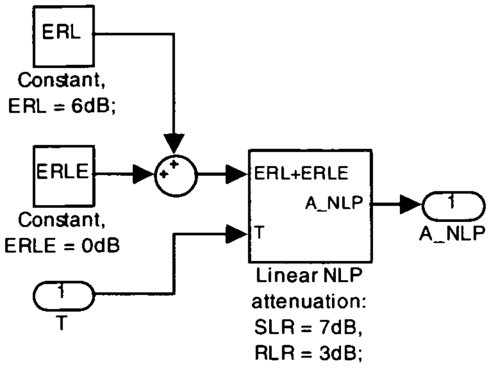

Figure A.12 – Simulation model of the ideal NLP control unit in Figure A.3.

The "misalign" measure, which is illustrated in Figure 4.5 and given by Equ. (4.25), has been obtained by modeling the following functions (Figure A.13).

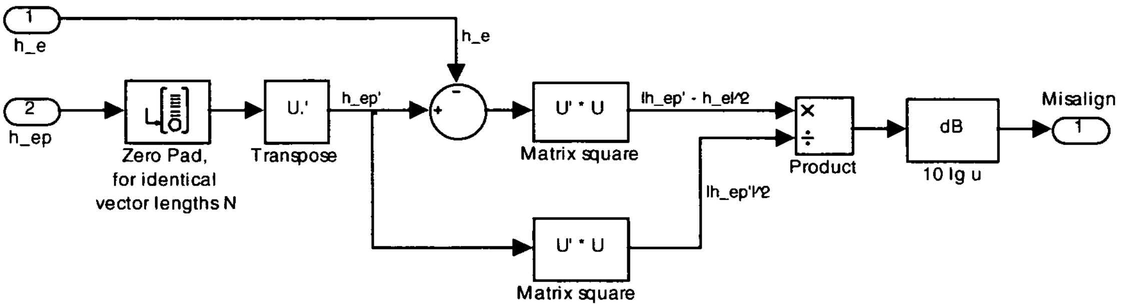

Figure A.13 – Simulation model for the "misalign" measure.

A.2.3 Echo-Path

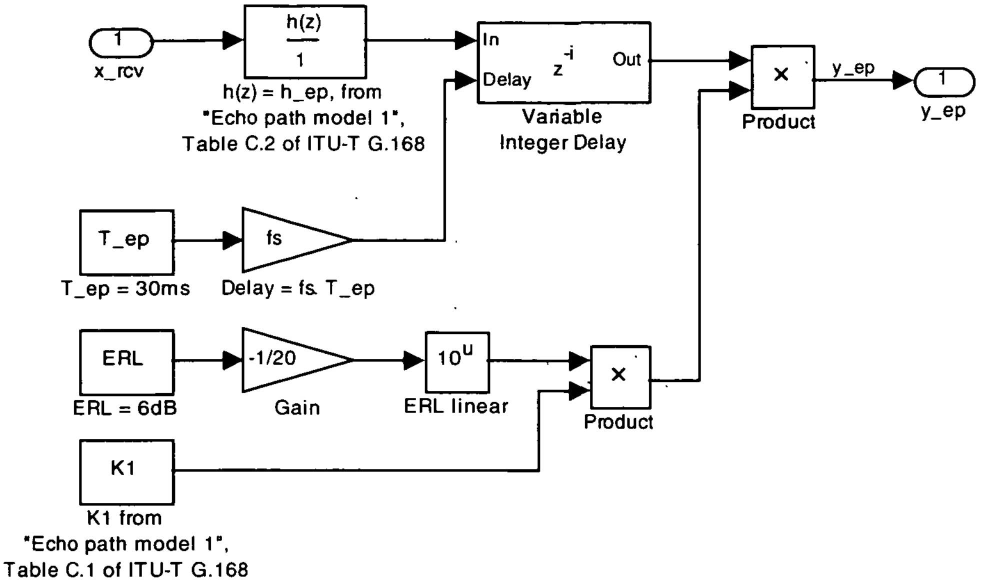

Figure A.14 – Simulation model for "echo path model 1" of ITU-T Recommendation G.168 [80].

A.3 References

A.3.1 Standard Double Talk

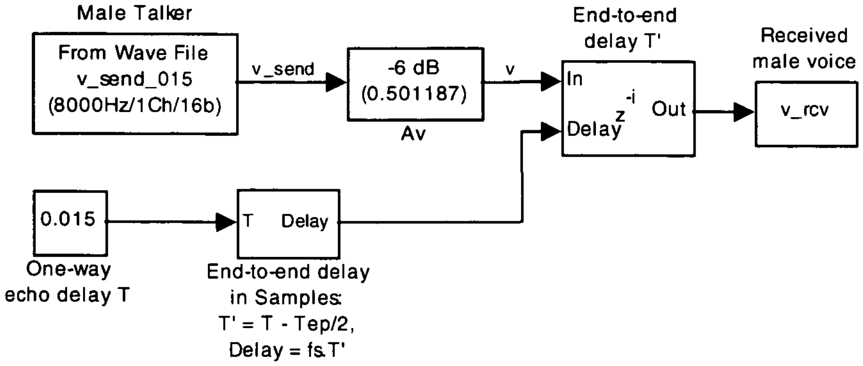

Figure A.15 – Simulation model for the "standard double talk" references; the corresponding parameters are listed in Table 4.5.

A.3.2 Far End Single Talk

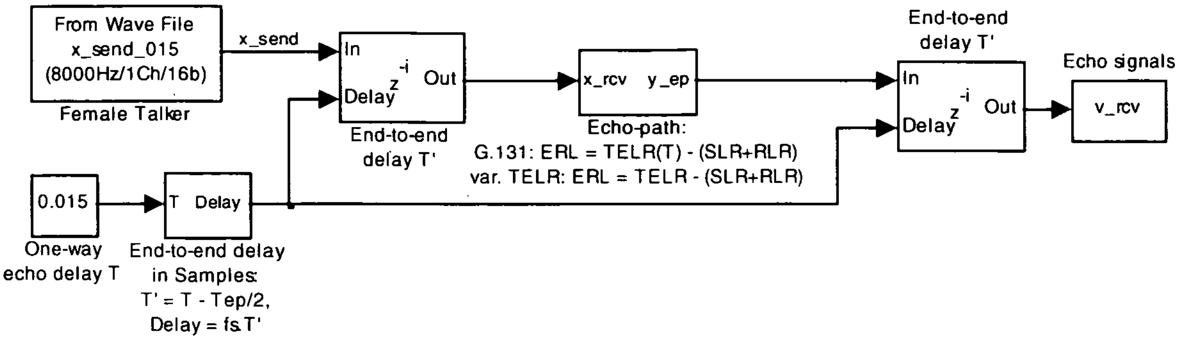

Figure A.15 – Simulation model for the "far end single talk" references. The echo-path is provided with two different sets of ERL values according to the parameters listed in Table 4.5.

Acronyms

ACR	Absolute Category Rating
AGC	Automatic Gain Control
AM	Amplitude Modulation
APA	Affine Projection Algorithm
API	Application Programming Interface
AQUILA	Adaptive resource control for Quality of service Using an IP-based Layered Architecture
ASL	Active Speech Level
AT&T	American Telephone and Telegraph Corporation
ATM	Asynchronous Transfer Mode
BICC	Bearer Independent Call Control
CCR	Comparison Category Rating
CD	Compact Disc
CNG	Comfort Noise Generation
CSS	Composite Source Signal
DCR	Degradation Category Rating
DDLC	Dynamic Driver Latency Compensation
DEC	Digital Echo Canceller
DECT	Digital Enhanced Cordless Telecommunications
DMOS	Degradation Mean Opinion Score
DSL	Digital Subscriber Line
DSP	Digital Signal Processor
DTD	Double Talk Detector
DTX	Discontinuous Transmission
ERL	Echo Return Loss
ERLE	Echo Return Loss Enhancement
ETSI	European Telecommunications Standard Institute
FAP	Fast Affine Projection
FEC	Forward Error Correction
FIR	Finite Impulse Response
FM	Frequency Modulation
FR	Frame Relay
FRLS	Fast Recursive Least-Squares
GEO	Geostationary Earth Orbit
GPS	Global Positioning System
GSM	Global System for Mobile communication
HATS	Head And Torso Simulator
HDLC	High level Data Link Control

ICMP	Internet Control Message Protocol
IETF	Internet Engineering Task Force
IN	Intelligent Network
IP	Internet Protocol
IPDV	Internet protocol Packet Delay Variation
IPPM	Internet Protocol Performance Metrics
ISC	International Softswitch Consortium
ISDN	Integrated Services Digital Networks
ISP	Internet Service Provider
IST	Information Society Technologies
ITU-T	International Telecommunication Union-Telecommunication Standardization Bureau
LAN	Local Area Network
LEO	Low Earth Orbit
LMS	Least Mean Square
LPF	Low-Pass Filter
LS	Least-Squares
MGCP	Media Gateway Control Protocol
MOS	Mean Opinion Score
MPLS	Multi Protocol Label Switching
MSC	Mobile Switching Center
MSF	Multiservice Switching Forum
NGN	Next Generation Network
NIMD	Non-Intrusive Measurement Device
NLMS	Normalized Least Mean Square
NLP	Non-Linear Processor
NTP	Network Time Protocol
OLR	Overall Loudness Rating
OPERA	Objective PERceptual Analyzer
OSI	Open System Interconnection
PACE	Perceived Annoyance Caused by Echoes
PAMS	Perceptual Analysis Measurement System
PC	Personal Computer
PCM	Pulse Code Modulation
PESQ	Perceptual Evaluation of Speech Quality
PLC	Packet Loss Concealment
PLMN	Public Land Mobile Network
PNLMS	Proportionate Normalized Least Mean Square
POTS	Plain Old Telephone Service
PSQM	Perceptual Speech Quality Measure
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RLR	Receiving Loudness Rating
RLS	Recursive Least-Squares

RSVP	Resource ReserVation Protocol
RTCP	Real-Time Transport Control Protocol
RTP	Real-Time Transport Protocol
SIP	Session Initiation Protocol
SIP-T	Session Initiation Protocol for Telephones
SLR	Sending Loudness Rating
SNR	Signal-to-Noise Ratio
SS7	Signaling System No. 7
STMR	SideTone Masking Rating
TCL	Terminal Coupling Loss
TCLw	Weighted Terminal Coupling Loss
TCP	Transmission Control Protocol
TDM	Time Division Multiplexing
TELR	Talker Echo Loudness Rating
THD	Total Harmonic Distortion
TIPHON	Telecommunication and Internet Protocol Harmonization Over Networks
UDP	User Datagram Protocol
UTC	Coordinated Universal Time
VAD	Voice Activity Detection
VLSI	Very Large Scale Integration
VoIP	Voice over Internet Protocol
VQT	Voice Quality Tester
WAN	Wide Are Network

References

[1] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay, *Advances in Network and Acoustic Echo Cancellation.* Springer, 2001.

[2] J. Davidson, J. Peters, *Voice over IP - Grundlagen.* Markt+Technik Verlag, 2000.

[3] G. Doblinger, *Signalprozessoren – Architekturen, Algorithmen, Anwendungen.* Schlembach Fachverlag Weil der Stadt, 2000.

[4] B. Douskalis, *IP Telephony, the integration of robust VoIP services.* Prentice Hall, 2000.

[5] E. Foth, *Handbuch IP-Telefonie.* Fossil-Verlag Köln, 2001.

[6] W. C. Hardy, *QoS measurement and evaluation of telecommunications quality of service*, Wiley, 2001.

[7] S. Haykin, *Adaptive Filter Theory.* Prentice Hall, Upper Saddle River, New Jersey, 3rd edition, 1996.

[8] O. Hersent, D. Gurle, J. P. Petit, *IP Telephony – Packet-based Multimedia Communication Systems.* Addison-Wesley 2000.

[9] R.-D. Köhler, *Voice over IP.* 1st edition, Bonn, mitp-Verlag , 2002.

[10] V. Kumar, M. Korpi, S. Sengodan, *IP Telephony with H.323: Architectures for Unified Networks and Integrated Services.* John Wiley & Sons, 2001.

[11] M. A. Miller, *Voice over IP - Strategies for the converged network.* IDG Books, 2000.

[12] D. Minoli, E. Minoli, *Delivering voice over Frame Relay and ATM.* John Wiley & Sons, 1998.

[13] D. Minoli, E. Minoli, *Delivering Voice over IP Networks.* Wiley Computer Publishing, 1998.

[14] N. J. Muller, *IP convergence: the next revolution in telecommunications.* Artech House, 2000.

[15] C. Norton, *Guide to Voice Fundamentals: From Analog to ATM.* First Edition, Nortel plc, 1997.

[16] D. J. Wright, *Voice over Packet Networks.* John Wiley & Sons, 2001.

[17] J. Berger, *Instrumentelle Verfahren zur Sprachqualitätsschätzung – Modelle auditiver Tests.* PhD thesis, Kiel, Shaker Verlag, 1998.

[18] P. Tschulik, *Konvergenz von kanal- und paketorientierten Sprachdiensten in öffentlichen Netzen.* PhD thesis, Vienna, 2002.

[19] S. Gustafsson, *Enhancement of Audio Signals by Combined Acoustic and Echo Cancellation and Noise Reduction.* PhD thesis, Wissenschaftsverlag Mainz, 1999.

[20] "Sprach-Echokompensator", Technische Spezifikation, Deutsche Telekom, TS-Nr. 0227/96, Feb. 1999.

[21] "Information: Digital-Echokompensator DEC, TED", S42023-A86-A1-4-18, Siemens ICN, 2001.

[22] "Transmission and Multiplexing (TM); Speech communication quality from mouth to ear for 3,1 kHz handset telephony across networks", ETSI Technical Report, July 1996.

[23] S. P. Applebaum, "Adaptive arrays", Syracuse Univ. Res. Corp., Rep. SPL-709, June 1964; IEEE Trans. Antennas Propagat., vol. 24, pp. 573-598, Sept. 1976.

[24] J. Benesty, T. Gänsler, "A robust fast recursive least squares adaptive algorithm", IEEE Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2001, vol. 6 , pp. 3785-3788, May 2001.

[25] J. Benesty, D. R. Morgan, J. H. Cho, "A new class of doubletalk detectors based on cross-correlation", IEEE Trans. Speech Audio Processing, vol. 8, pp. 168-172, Mar. 2000.

[26] W. Brandstätter, F. Kettler: "Perception oriented, delay-controlled echo cancellation in IP based telephone networks", International Workshop on Acoustic Echo and Noise Control (IWAENC), Kyoto, pp. 199-202, Sept. 2003.

[27] W. Brandstätter, F. Kettler: "Delay-controlled echo cancellation in IP based telephone networks", International Conference on Computer, Communication and Control Technologies (CCCT), Orlando, vol. 5, pp. 164-168, July 2003.

[28] W. Brandstätter, B. Handl, H. Jammernegg, W. Müllner, P. Tschulik: "A New Approach of Echo Cancellation Design in IP based Telephone Networks", DAGA, Aachen, pp. 786-787, March 2003.

[29] W. Brandstätter, F. Kettler: "Laufzeitgesteuerte Echokompensation in IP basierenden Telefonnetzen", Elektronische Sprachsignalverarbeitung (ESSV), Karlsruhe, pp. 154-161, Sept. 2003.

[30] W. Brandstätter, F. Kettler: "Ein neuer Ansatz für Echokompensation in IP basierenden Telefonnetzen", Mikroelektronik Tagung, Vienna, pp. 335-341, Oct. 2003.

[31] D. L. Duttweiler, "A twelve-channel digital echo canceller", IEEE Trans. Commun., vol. 26, pp. 647-653, May 1978.

[32] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancellers", IEEE Trans. Speech Audio Processing, vol. 8, pp. 508-518, Sept. 2000.

[33] E. Eweda, "Comparison of RLS, LMS, and sign algorithms for tracking randomly time-varying channels", IEEE Trans. Signal Processing, vol. 42, pp. 2937-2944, Nov. 1994.

[34] T. Gänsler, J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 215-218, Oct. 1999.

[35] T. Gänsler, J. Benesty, S. L. Gay, M. M. Sondhi, "A robust proportionate affine projection algorithm for network echo cancellation", IEEE Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2000, vol. 2, pp. II793-II796, June 2000.

[36] T. Gänsler, M. Hansson, C.-J. Ivarsson, G. Salomonsson, "A double-talk detector based on coherence", IEEE Trans. Comm., vol. 44, pp. 1421-1427, Nov. 1996.

[37] S. L. Gay, "An efficient, fast converging adaptive filter for network echo cancella-

tion", Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems and Computers, vol. 1, pp. 394-398, Nov. 1998.

[38] S. L. Gay, M. M. Sondhi, J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation", IEEE Trans. on Speech and Audio Processing, vol. 8, no. 6, pp. 656-663, Nov. 2000.

[39] W. Jiang, H. Schulzrinne, "QoS measurement of internet real-time multimedia services", Technical Report CUCS-015-99m, Columbia Univ., New York, Dec. 1999.

[40] F. Kettler, H. W. Gierlich, H. Kullmann, "Sprachechokompensatoren im Telefonnetz: Qualitätsbestimmende Parameter", DAGA, Kiel, 1997.

[41] S. Marple, "Efficient least-squares FIR system identification", IEEE Trans. Acoust., Speech, Signal Processing, vol. 29, pp. 62-73, 1981.

[42] D. Messerschmitt, "Echo cancellation in speech and data transmission", IEEE J. Selected Areas Commun., vol. 2, pp. 283-297, Mar. 1984.

[43] D. Mills, P.H. Kamp, "The nanokernel", Proc. Precision time and time interval (PTTI) applications and planning meeting, Reston, pp. 423-430, Nov. 2000.

[44] D. R. Morgan, S. G. Kratzer, "On a class of computationally efficient, rapidly converging, generalized NLMS algorithms", IEEE Signal Processing Lett., vol. 3, pp. 245-247, Aug. 1996.

[45] K. Ochiai, T. Araseki, T. Ogihara, "Echo canceller with two echo path models", IEEE Trans. Commun., vol. 25, pp. 589-595, June 1977.

[46] K. Ozeki, T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties", Electron Commun. Japan, vol. 67-A, pp. 19-27, 1984.

[47] M. M. Sondhi, D. A. Berkley, "Silencing echoes on the telephone network", IEEE Proc., vol. 68, pp. 948-963, Aug. 1980.

[48] M. M. Sondhi, "An adaptive echo canceller", Bell Syst. Tech. J., vol. 46, pp. 497-511, March 1967.

[49] M. M. Sondhi, "Echo canceller", U.S. Patent 3,499,999, Mar. 10, 1970 (filed Oct. 31, 1966).

[50] O. Tanrikule, K. Dogancay, "Adaptive filtering algorithms with selective partial updates", IEEE Trans. on Circuits and Systems, vol. 48, no. 8, pp. 762-769, Aug. 2001.

[51] O. Tanrikule, K. Dogancay, "Selective-partial-update proportionate normalized least-mean-squares algorithm for network echo cancellation", IEEE, pp. II1889-II1892, 2002

[52] B. Widrow, M. E. Hoff Jr., "Adaptive switching circuits", in IRE Wescon Conc. Rec., part 4, pp. 96-104, 1960.

[53] H. Ye, B.-X. Wu, "A new double-talk detection algorithm based on the orthogonality theorem", IEEE Trans. Commun. vol. 39, pp. 1542-1545, Nov. 1991.

[54] J. Gozdecki, A. Jajszczyk, R. Stankiewicz, "Quality of service terminology in IP networks", IEEE Communications Magazine, vol. 41, issue 3, pp. 153-159, March 2003.

[55] P. Denisowski, "How does it sound?", IEEE Spectrum, vol. 38, issue 2, pp. 60-64, Feb. 2001.

[56] M. Hassan, A. Nayandoro, M. Atiquzzaman, “Internet telephony: services, technical challenges, and products“, IEEE Communications Magazine, vol. 38, issue 4, pp. 96-103, April 2000.

[57] R. Ramjee, J. Kurose, D. Towsley, “Adaptive playout mechanisms for packetized audio applications in wide-area networks“, 13th Proceedings of IEEE INFOCOM '94, Networking for Global Communications, pp. 680-688, June 1994.

[58] R. J. B. Reynolds, A. W. Rix, “Quality VoIP - an engineering challenge“, BT Technology Journal, vol. 19, issue 2, pp. 23-32, April 2001.

[59] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs“, IEEE Proc. Int. Conf. Acoustics, Speech, Signal Processing, vol. 2, pp. 749-752, 2001.

[60] A. W. Rix, M. P. Hollier, “The perceptual analysis measurement system for robust end-to-end speech quality assessment“, IEEE Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1515-1518, June 2000.

[61] F. Kettler, H. W. Gierlich, E. Diedrich, J. Berger, “Echobeurteilung beim Abhören von Kunstkopfaufnahmen im Vergleich zum aktiven Sprechen“, DAGA, Oldenburg, 2001.

[62] F. Kettler, H. W. Gierlich, E. Diedrich, “Echo and speech level variations during double talk influencing hands-free telephone transmission quality“, International Workshop on Acoustic Echo and Noise Control, Pocona Manor, 1999.

[63] H. W. Gierlich, F. Kettler, “Background noise transmission and comfort noise insertion: The influence of signal processing on speech-quality in complex telecommunication scenarios“, International Workshop on Acoustic Echo and Noise Control, Darmstadt, Sept. 2001.

[64] ITU-T Rec. E.800, “Terms and definitions related to quality of service and network performance including dependability“, Aug. 1994.

[65] ITU-T Rec. G.107, “The E-model, a computational model for use in transmission planning“, March 2003.

[66] ITU-T Rec. G.108, “Application of the E-model: A planning guide“, Sept. 1999.

[67] ITU-T Rec. G.108.2, “Transmission planning aspects of echo cancellers“, Jan. 2003.

[68] ITU-T Rec. G.109, “Definition of categories of speech transmission quality“, Sept. 1999.

[69] ITU-T Rec. G.111, “Loudness ratings (LRs) in an international connection“, March 1993.

[70] ITU-T Rec. G.114, “One-way transmission time“, May 2003.

[71] ITU-T Rec. G.116, “Transmission performance objectives applicable to end-to-end international connections“, Sept. 1999.

[72] ITU-T Rec. G.121, “Loudness ratings (LRs) of national systems“, March 1993.

[73] ITU-T Rec. G.122, “Influence of national systems on stability and talker echo in international connections“, March 1993.

[74] ITU-T Rec. G.131, “Stability and echo“, 1988.

[75] ITU-T Rec. G.131, “Control of talker echo“, August 1996.

[76] ITU-T Rec. G.131 App. II, “Relation between echo disturbances under single talk and double talk conditions (evaluated for one-way transmission time of 100 ms), September 1999.

[77] ITU-T Rec. G.164, “Echo suppressors”, Nov. 1988.

[78] ITU-T Rec. G.165, “Echo cancellers“, March 1993.

[79] ITU-T Rec. G.167, “Acoustic echo controllers“, March 1993.

[80] ITU-T Rec. G.168, “Digital network echo canceller“, June 2002.

[81] ITU-T Rec. G.223, “Assumptions for the calculation of noise on hypothetical reference circuits for telephony“, Nov. 1988.

[82] ITU-T Rec. G.711, “Pulse code modulation (PCM) of voice frequencies“, Nov. 1988.

[83] ITU-T Rec. G.711 App. II, “A comfort noise payload definition for ITU-T G.711 use in packet-based multimedia communication systems”, Feb. 2000.

[84] ITU-T Rec. G.723.1, “Speech coders: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s“, March 1996.

[85] ITU-T Rec. G.723.1, Annex A, “Silence compression scheme“, Nov. 1996.

[86] ITU-T Rec. G.726, “40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM)”, Dec. 1990.

[87] ITU-T Rec. G.729, “Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)“, March 1996.

[88] ITU-T Rec. G.729, Annex B, “A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70“, Oct. 1996.

[89] ITU-T Rec. H.323, “Packet-based multimedia communications systems”, July 2003.

[90] ITU-T Rec. H.248.1, “Gateway control protocol: version 2“, May 2002.

[91] ITU-T Rec. I.113, “Vocabulary of terms for broadband aspects of ISDN”, June 1997.

[92] ITU-T Rec. P.50, “Artificial voices“, Sept. 1999.

[93] ITU-T Rec. P.56, “Objective measurement of active speech level“, March 1993.

[94] ITU-T Rec. P.57, “Artificial Ears”, July 2002.

[95] ITU-T Rec. P.58, “Head and torso simulator for telephonometry“, Aug. 1996.

[96] ITU-T Rec. P.59, “Artificial conversational speech”, March 1993.

[97] ITU-T Rec. P.79, “Calculation of loudness ratings for telephone sets“, Sept. 1999.

[98] ITU-T Rec. P.310, “Transmission characteristics for telephone band (300-3400 Hz) digital telephones“, March 2003.

[99] ITU-T Rec. P.340, “Transmission characteristics and speech quality parameters of hands-free terminals“, May 2000.

[100] ITU-T Rec. P.501, “Test signals for use in telephonometry“, May 2000.

[101] ITU-T Rec. P.502, “Objective test methods for speech communication systems using complex test signals“, May 2000.

[102] ITU-T Rec. P.561, “In-service, non-intrusive measurement device – voice service measurements“, Feb. 1996.

[103] ITU-T Rec. P.562, “Analysis and interpretation of INMD voice-service measurements“, May 2000.

[104] ITU-T Rec. P.800, “Methods for subjective determination of transmission quality“, Aug. 1996.

[105] ITU-T Rec. P.810, “Modulated noise reference unit (MNRU)“, Feb. 1996.

[106] ITU-T Rec. P.830, “Subjective performance assessment of telephone-band and wideband digital codecs“, Feb. 1996.

[107] ITU-T Rec. P.831, “Subjective performance evaluation of network echo cancellers“, Dec. 1998.

[108] ITU-T Rec. P.832, “Subjective performance evaluation of hands free terminals“, May 2000.

[109] ITU-T Rec. P.861, “Objective quality measurement of telephone-band (300-3400 Hz) speech codecs“, Feb. 1998.

[110] ITU-T Rec. P.862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs“, Feb. 2001.

[111] ITU-T Rec. Q.1902.1, “Bearer independent call control protocol (capability set 2): Functional description”, July 2001.

[112] ITU-T Rec. V.8, “Procedures for starting sessions of data transmission over the public switched telephone network”, Nov. 2000.

[113] ITU-T Rec. V.21, “300 bits per second duplex modem standardized for use in the general switched telephone network”, Nov. 1988.

[114] ITU-T Rec. V.90, “A digital modem and analogue modem pair for use on the public switched telephone network (PSTN) at data signalling rates of up to 56 000 bit/s downstream and up to 33 600 bit/s upstream”, Sept. 1998.

[115] ITU-T Rec. Y.1241, “Support of IP based services using IP transfer capabilities” March 2001.

[116] ITU-T Rec. Y.1540, “Internet protocol data communication service – IP packet transfer and availability performance parameters”, Dec. 2002.

[117] ITU-T Rec. Y.1541, “Network performance objectives for IP-based services”, May 2002.

[118] ITU-T Contribution, “Conversational tests with speech echo cancellers – description of test procedures and results”, ITU-T rapporteurs meeting, Jerusalem, Oct. 1996.

[119] ITU-T Contribution COM 12-16, “Auditory judgement of echo: Talking and listening test in comparison to third party listening test”, Dec. 2000.

[120] ITU-T Contribution COM 12-103, “Relation between echo disturbances under single talk and double talk conditions (evaluated for one-way transmission time of 100 ms”, July 1999.

[121] ITU-T Study Group 12 Contribution 10, COM 12-10-E, “Proposed draft recommendation on the perceptual echo and sidetone quality measure (PESQM), an objective method for talking quality assessment”, Nov. 2000.

[122] ITU-T, ”Handbook on Telephonometry”, 2nd edition, Geneva, 1992.

[123] IETF RFC 791, “Internet protocol: DARPA Internet program protocol specification”, Sept. 1981.

[124] IETF RFC 793, “Transmission Control Protocol”, Sept. 1981.

[125] IETF RFC 768, “User Datagram Protocol”, Aug. 1980.

[126] IETF RFC 1305, “Network time protocol (NTP) version 3: specification, implementation and analysis“, March 1992.

[127] IETF RFC 2427, “Multiprotocol interconnect over Frame Relay”, Sept. 1998.

[128] IETF RFC 1633, “Integrated services in the Internet architecture: an overview”, June 1994.

[129] IETF RFC 1889, “RTP: A transport protocol for real-time applications“, Jan. 1996.

[130] IETF RFC 1890, “RTP profile for audio and video conferences with minimal control“, Jan. 1996.

[131] IETF RFC 2030, “Simple network time protocol (SNTP) version 4 for IPv4, IPv6 and OSI“, Oct. 1996.

[132] IETF RFC 2205, “Resource reservation protocol (RSVP) -- version 1 functional specification“, Sept. 1997.

[133] IETF RFC 2330, “Framework for IP performance metrics“, May 1998.

[134] IETF RFC 2386, “A framework for QoS-based routing in the Internet“, Aug. 1998.

[135] IETF RFC 2475, “An architecture for differentiated services“, Dec. 1998.

[136] IETF RFC 2508, “Compressing IP/UDP/RTP headers for low-speed serial links“, Feb. 1999.

[137] IETF RFC 2679, “A one-way delay metric for IPPM“, Sept. 1999.

[138] IETF RFC 2680, “A one-way packet loss metric for IPPM”, Sept. 1999.

[139] IETF RFC 2681, “A round-trip delay metric for IPPM”, Sept. 1999.

[140] IETF RFC 2705, “Media gateway control protocol (MGCP) version 1.0“, Oct. 1999.

[141] IETF RFC 3015, “Megaco protocol version 1.0“, Nov. 2000.

[142] IETF RFC 3031, “Multiprotocol Label Switching Architecture”, Jan. 2001.

[143] IETF RFC 3261, “SIP: Session initiation protocol”, June 2002.

[144] IETF RFC 3372, “Session initiation protocol (SIP) for telephones (SIP-T): Context and architectures”, Sept. 2002.

[145] IETF RFC 3393, “IP packet delay variation metric for IP performance metrics (IPPM)“, Nov. 2002.

[146] ETSI EG 201 377-1, version 1.2.1, “Speech processing, transmission and quality aspects (STQ); Specification and measurement of speech transmission quality; Part 1: Introduction to objective comparison measurement methods for one-way speech quality across networks“, Dec. 2002.

[147] ETSI ETR 003, 2nd edition, “Network aspects (NA); General aspects of quality of service (QoS) and network performance (NP)”, Oct. 1994.

[148] ETSI ETR 250, “Transmission and multiplexing (TM); Speech communication quality from mouth to ear for 3,1 kHz handset telephony across networks“, July 1996.

[149] ETSI GTS 06.10, V3.2.0, “European digital cellular telecommunications system (phase 1); GSM full rate speech transcoding (GSM 06.10)”, Jan. 1995

[150] ETSI TBR 8 edition 2, “Integrated services digital network (ISDN), telephony 3,1 kHz teleservice”, Oct. 1998.

[151] ETSI TBR 21, “Terminal equipment (TE); Attachment requirements for pan Euro-

pean approval for connection to the public switched telephone networks (PSTNs) of TE (excluding TE supporting the voice telephony service) in which network addressing, if provided, is by means of dual tone multi frequency (DTMF) signalling", Jan. 1998.

[152] ETSI TS 101 329-1, version 3.1.2, "Telecommunications and Internet Protocol harmonization over networks (TIPHON) release 3; End-to-end quality of service in TIPHON systems; Part 1: General aspects of quality of service (QoS)", Jan. 2002.

[153] ETSI TS 101 329-2, version 2.1.3, "Telecommunications and Internet Protocol harmonization over networks (TIPHON) release 3; End-to-end quality of service in TIPHON systems; Part 2: Definition of speech quality of service (QoS) classes", Jan. 2002.

[154] ETSI TS 101 329-5, version 1.1.2: "Telecommunications and Internet Protocol harmonization over networks (TIPHON) release 3; End-to-end quality of service in TIPHON systems; Part 5: Quality of service (QoS) measurement methodologies", Jan. 2002.

[155] F. Kettler et al., "Anonymized Test Report, 2nd ETSI Speech Quality Test Event for Voice over IP", April 2001.

[156] ANSI TIA/EIA-810-A, "Transmission requirements for narrowband voice over IP and voice over PCM digital wireline telephones", Dec. 2000.

[157] ANSI TIA/EIA/TSB116, "Telecommunications, IP telephony equipment, voice quality recommendations for IP telephony", March 2001.

Cybergraphy

[158] International Softswitch Consortium, http://www.softswitch.org/.

[159] "Reference architecture", version 1.2, International Softswitch Consortium, http://www.softswitch.org/publications/, June 2002.

[160] Multiservice Switching Forum, http://msforum.org/.

[161] ETSI Project Telecommunication and Internet Protocol Harmonization over Networks, http://portal.etsi.org/portal_common/ home.asp?tbkey1= TIPHON.

[162] ETSI Speech Transmission Quality (STQ), http://portal.etsi.org/portal_common/home.asp?tbkey1=STQ

[163] VocalTec, http://www.vocaltec.com/.

[164] Network Time Protocol (NTP) project, http://www.ntp.org/.

[165] IETF IP performance metrics (IPPM) working group, http://www.ietf.org/html.charters/ippm-charter.html.

[166] Adaptive Resource Control for QoS Using an IP-based Layered Architecture (AQUILA) project, IST-1999-10077, http://www.ist-aquila.org/.

[167] Meinberg Radio Clocks, http://www.meinberg.de/.

[168] Agilent Voice Quality Tester (VQT), http://www.agilent.com/comms/voicequality/.

[169] SwissQual NetQual, http://www.swissqual.com/html/netqualpage.htm

[170] Opticom Objective Perceptual Signal Quality Analyzer (OPERA), http://www.opticom.de/3_products/3opera-set.html.

[171] Internet Engineering Task Force (IETF), http://www.ietf.org/.

[172] International Telecommunication Union (ITU) Telecommunication Standardization Sector (ITU-T), http://www.itu.int/ITU-T/.

[173] International Telecommunication Union (ITU) Telecommunication Standardization Sector (ITU-T) Study Group 12, http://www.itu.int/ITU-T/studygroups /com12/.

Curriculum vitae

Dipl.-Ing. Wolfgang Brandstätter

19. Februar 1974	geboren in Grieskirchen
1980 – 1984	Volksschule in Offenhausen
1984 – 1988	Bundesrealgymnasium Wallererstraße in Wels
1988 – 1993	Höhere Technische Bundeslehranstalt für Elektrotechnik in Wels Matura mit Auszeichnung
1993 – 1994	Grundwehrdienst in Wels
Oktober 1994	Beginn des Studiums der Elektrotechnik, Studienzweig Automatisierungs- und Regelungstechnik
April 2000	Abschluss des Studiums mit Auszeichnung, Diplomarbeit *Analyse und Optimierung von Regelkonzepten für geschaltete Leistungsverstärker in Multizellenstruktur*
Mai 2000 – September 2000	Firma Konrad M&R – Messtechniker für Revisionsarbeiten in deutschen Kernkraftwerken
Oktober 2000 – März 2003	Beginn als wissenschaftlicher Mitarbeiter am Institut für Elektrische Mess- und Schaltungstechnik der Technischen Universität Wien Forschungsprojekte mit Partnern aus der Industrie Forschungsschwerpunkt auf Next Generation Networks Architekturen, Voice over IP, subjektive und objektive Messverfahren zur Bestimmung der Sprachqualität
September 2003 – Februar 2004	Honorarlehrer für Elektrotechnik-Kurse am Berufsförderungsinstitut Wien