

ITWissen
 Das große Online-Lexikon
für Informationstechnologie

SPRACHQUALITÄT

KLAUS LIPINSKI /Hrsg.]

Inhalt

ACELP, algebraic code excited linear prediction

AI, articulation index

ALcons, articulation loss of consonants

BER, bit error rate

CAC, call admission control

CELP, code excited linear prediction

CS-ACELP, conjugate structure algebraic code excited linear prediction

E-Modell

EFR, enhanced full rate

Fernsprechen,

G.107

G.114

Hörcharakteristik

Hörschwelle

LD-CELP, low delay, code excited linear prediction

MOS, mean opinion score

Paketverlust

Paketverlustrate

PAMS, perceptual analysis measurement system

PESQ, perceptual evaluation of speech quality

POTS, plain old telephone service

PSQM, perceptual speech quality measurement

R, transmission quality rating

RASTI, rapid speech transmission index

Rauschen

SBR, spectral band replication

Signal-Rausch-Verhältnis, S/N

SII, speech intelligibility index

Sprachcodec

Sprache

Sprachfrequenz

Sprachkommunikation

Sprachkompression

Sprachqualität

Sprachverständlichkeit

Sprachübertragungsindex

Verständlichkeit

Verzögerung

VG, voice grade

Impressum:

Herausgeber: Klaus Lipinski

Sprachqualität

Copyright 2010

DATAKOM-Buchverlag GmbH

84378 Dietersburg

Alle Rechte vorbehalten.

Keine Haftung für die angegebenen Informationen. Das E-Book ist urheberrechtlich geschützt und darf nicht auf fremden Websites ins Internet oder in Intranets gestellt werden.

Produziert von Media-Schmid

www.media-schmid.de

ACELP, algebraic code excited linear prediction

Algebraic Code Excited Linear Prediction (ACELP) ist eine *Sprachkompression* für Bandbreiten zwischen 2,4 kbit/s und 8 kbit/s. ACELP ist eine verbesserte Kompressionstechnik, die in *Sprachcodecs* mit Code Excited Linear Prediction (CELP) und im Audiocodecs G.723.1 von UMTS und dem *EFR-Verfahren* von GSM eingesetzt wird.

Die durch den *MOS-Wert* ermittelte *Sprachqualität* liegt bei 3,9.

AI, articulation index

Es gibt mehrere Messverfahren zur Bewertung der *Sprachverständlichkeit* in beschallten Räumen und Sprachkommunikations-Systemen. Der Articulation Index (AI) ist eines der ersten Verfahren und wurde von den Bell Telephone Labs bereits in den vierziger Jahren entwickelt.

Der Articulation Index geht davon aus, dass sich die Verständlichkeit in Sprachkommunikations-Systemen aus der Summe der Verständlichkeit einzelner Frequenzbänder des Sprachbereichs zusammensetzt. Deshalb wird im AI-Index der Sprachfrequenzbereich in viele Frequenzbänder, Oktaven oder Terzen unterteilt und von jedem einzelnen werden die Eigenschaften ermittelt. Darüber hinaus wird für jedes Frequenzband auch das *Signal-Rausch-Verhältnis* ermittelt und gewichtet. Die Summe der Übertragungseigenschaften ist der Articulation Index.

Der AI-Index kann wie der *Speech Transmission Index (STI)* Werte zwischen "0" und "1" annehmen. "0" steht wie beim STI-Index für eine nicht akzeptierbare *Sprachqualität*, "1" für eine exzellente.

ALcons, articulation loss of consonants

ALcons ist wie der *Speech Transmission Index (STI)* ein Bewertungsverfahren für die *Sprachverständlichkeit* in beschallten Räumen und Sälen. Wie aus der Bezeichnung Alcons hervorgeht, handelt es sich um den Artikulationsverlust bei Konsonanten. Da die Konsonanten maßgeblich sind für die *Sprachverständlichkeit*, stützt sich die Alcons-Bewertung auf den prozentualen Anteil der akustisch nicht korrekt verstandenen Konsonanten. Diese werden in beschallten Räumen durch Nachhall, Reflexionen und Störungen die Sprachverständlichkeit beeinträchtigt.

Der ALcons-Wert wird in Prozent angegeben und ist von den genannten Faktoren sowie von der

Richtcharakteristik der Lautsprecherboxen, der Strahlbündelung, dem Abstand der Zuhörer, der Aufbau des Beschallungsraums, dessen Wand- und Bodenbeschaffenheit, den Reflexion und von vielem mehr abhängig. Die genannten Faktoren werden bei Alcons in Relation zueinander gesetzt und daraus werden die prozentualen Alcons-Werte ermittelt. Diese können zwischen 100 % und 0 % liegen, wobei der 100-%-Wert die schlechteste Sprachverständlichkeit repräsentiert, der 0-%-Wert die beste.

STI-Index SII-Index	Sprachverständlichkeit	Alcons
0 bis 0,3	Nicht akzeptierbar, unverständlich	100 % bis 33 %
0,3 bis 0,45	Schlecht	33 % bis 15 %
0,45 bis 0,6	Genügend	15 % bis 7 %
0,6 bis 0,75	Gut	7 % bis 3 %
0,75 bis 1,0	Ausgezeichnet	3 % bis 0 %

Bewertung der Sprachverständlichkeit nach dem STI-, SII-Index und ALcons

BER, bit error rate BFR, Bitfehlerrate

$$BER = \frac{B_f}{B_g} = \frac{B_f}{\dot{U}_r \times T_{int}}$$

B_f, fehlerhafte Bits
B_g, Gesamtbitzahl
Ü_r, Übertragungsrate (bit/s)
T_{int}, Zeitintervall der Übertragung

Bestimmung der Bitfehlerrate

Die Bitfehlerrate (BER) ist das Verhältnis der Anzahl der binären Signalelemente, die bei der Übertragung verfälscht wurden, zur Gesamtzahl der ausgesendeten binären Signalelemente. Eine Bitfehlerhäufigkeit von 1 bedeutet, dass jedes Bit falsch ist. Eine Fehlerrate von 6×10^{-6} bedeutet, dass durchschnittlich 6 Bits falsch sein können, wenn 1 Million Bits übertragen werden.

Für die verschiedenen Lokalen Netze schreibt der Standard unterschiedliche Werte für die Bitfehlerrate vor: Für Ethernet fordert der Standard einen BER-Wert von 10^{-8} , für Token Ring von 10^{-9} und für FDDI $2,5 \times 10^{-12}$. Das bedeutet, dass bei FDDI ein fehlerhaftes Bit auf 400 Milliarden übertragene Bits entfällt. In der *Sprachkommunikation* macht sich die Bitfehlerrate unterschiedlich störend bemerkbar. Eine Bitfehlerrate von 10^{-2} macht sich als störendes Prasseln bemerkbar, bei 10^{-3} löst sich das Prasseln in einer dichten Folge von Knackgeräuschen auf, die bei einer Bitfehlerrate von 10^{-4} in einzelne Knackgeräusche übergeht. Bei 10^{-5} sind nur noch einzelne Knackgeräusche hörbar und Bitfehlerraten von 10^{-6} beeinträchtigen die Sprachübertragung überhaupt nicht.

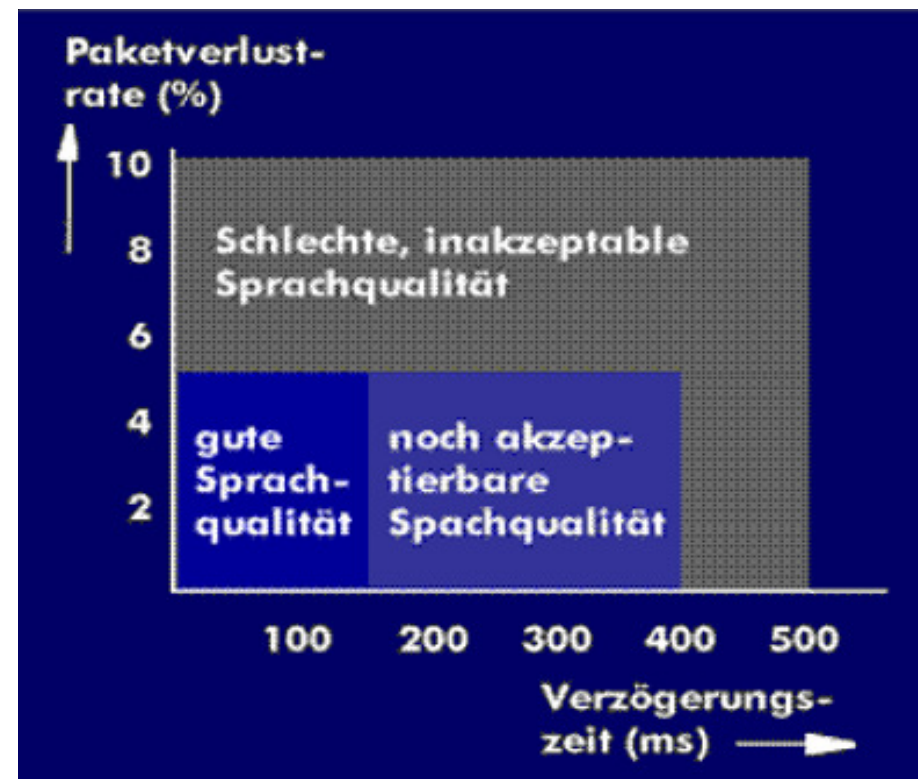
CAC, call admission control

Call Admission Control (CAC) ist ein Verfahren zur Verbesserung der *Sprachqualität* bei der Übertragung von *Sprache* und anderen Echtzeit-Anwendungen in IP-Netzen. Mit der CAC- Technik soll die Sprachqualität der Internettelefonie und von VoIP gesichert werden. Call Admission Control verhindert die Überbelastung der vorhandenen Bandbreite und die damit in Zusammenhang stehende Beeinträchtigung der *Sprachverständlichkeit*.

Beim CAC-Verfahren werden IP-basierte Sprachübertragungen gegen Überlastung geschützt indem der Aufbau von weiteren Sprachverbindungen, die die Sprachqualität verschlechtern, verhindert wird. Wird nämlich die zulässige Anzahl an Sprachverbindungen überschritten, müssen die vorhandenen

Sprachverbindungen den zusätzlichen Sprachverkehr aufnehmen, was zu *Paketverlusten* führt und die Sprachqualität aller anderen Sprachverbindungen beeinträchtigt.

CELP, code excited linear prediction



ITU-Empfehlung G.114 für Verzögerungszeiten bei Sprachübertragungen

Code Excited Linear Prediction (CELP) ist ein hybrides Verfahren der *Sprachkompression*, das die Vorteile der Signalformcodierung mit denen der parametrischen Codierung vereint. Es kombiniert Pulscodemodulation (PCM) mit Parametric Stereo (PS), wie der LPC-Codierung. CELP zeichnet sich durch eine gute *Sprachqualität* aus, vergleichbar mit PCM, hat allerdings eine geringere Datenrate im Vergleich zu PCM oder ADPCM. Ein hybrider Vocoder basierend auf CELP hat bedingt durch die doppelte Codierung eine wesentlich höhere Komplexität.

Die Datenraten von CELP sind in 200-bit/s-Stufen

skalierbar zwischen 3,85 kbit/s und 23,8 kbit/s, bei Abtastraten zwischen 8 kHz und 16 kHz. Mit diesen Datenraten kann Musik nur in verminderter Qualität übertragen werden.

Neben dem normalen CELP-Verfahren gibt es von der ITU-T spezifizierte Varianten mit geringeren *Verzögerungszeiten* und Übertragungsraten, das so genannte »Low Delay CELP (*LD-CELP*)«, das in der ITU-T-Empfehlung G.728 spezifiziert ist und das »Conjugate Structure *ACELP (CS-ACELP)*« aus G.729. G.728 erreicht bei einer Übertragungsrate von 16 kbit/s einen *MOS-Wert* für die Sprachqualität von 4,0 und zeichnet sich durch eine Signalverzögerungszeit aus, die durch das Codieren und Decodieren entsteht, die bei 0,625 ms liegt. Bei dem Standard G.729 wird vor dem Codieren ein Vergleich des Sprachsignals mit dem Modell durchgeführt. Dieses aufwändige Verfahren benötigt für die Übertragung nur die halbe Übertragungsrate gegenüber LD-CELP.

Für das in MPEG-4 eingesetzte CELP gibt es zwei Abtastfrequenzen von 8 kHz und 16 kHz. Der Betrieb mit der niedrigeren Abtastrate wird als NB-CELP (Narrowband) bezeichnet, das mit 16 kHz als WB-CELP (Wideband).

CS-ACELP, conjugate structure algebraic code excited linear prediction

Bei *CS-ACELP* (Conjugate Structure Algebraic Code Excited Linear Prediction) handelt es sich um einen Algorithmus zur *Sprachkompression* in Übertragungsleitungen mit 8 kbit/s Übertragung. Der Algorithmus, der im ITU-Standard G.729 (A) beschrieben ist, regelt auch die Sprechpausenunterdrückung. Während dieser Pausen wird die zur Verfügung stehende Bandbreite einer anderen Netzanwendung genutzt. Die durch den *MOS-Wert* ermittelte *Sprachqualität* ist als sehr gut zu bezeichnen und liegt bei 4,2.

ACELP ist der vom Frame Relay Forum vorgeschlagene Algorithmus für die Übertragung von Voice over Frame Relay (VoFR).

E-Modell

Das E-Modell ist ein von der ITU unter *G.107* standardisiertes Berechnungsmodell für die Bewertung der *Sprachqualität* in Übertragungssystemen. Mit diesem Modell, das bei der Planung und Simulation von Netzen

R-Faktor	Bewertung der Sprachqualität
100	Exzellente, gut verständlich
G.107 → 94 Default-Wert	
80	Gut, Sprache kann verstanden werden
60	Ordentlich, beim Zuhören ist eine gewisse Konzentration erforderlich
50	Mäßig, man kann die Sprache nur mit hoher Konzentration verstehen
50 bis 0	Mangelhaft, es ist keine Verständigung möglich

benutzt wird, wird der objektive *R-Faktor*, der eine Aussage über die Sprachqualität gibt, in Abhängigkeit von verschiedenen Übertragungseinflüssen ermittelt.

Der dabei benutzte Ansatz berücksichtigt nicht einen einzelnen, individuellen Übertragungsparameter, sondern alle Parameter, die Einfluss auf die Übertragungsqualität haben mit ihrer gegenseitigen Abhängigkeit. Zu den die Sprachqualität und die beiden Werte beeinflussenden Größen gehören das *Rauschen*, das *Signal-Rausch-Verhältnis*, Verzögerungen,

Objektive Sprachbewertung durch den R-Faktor

Jitter, Latenzzeiten, *Echos* und *Paketverluste*.

Das E-Modell zeichnet sich dadurch aus, dass es die Parameter der Simulation direkt als Eingangsparameter für die Vorhersage der Sprachqualität benutzt.

EFR, enhanced full rate *EFR-Verfahren*

Enhanced Full Rate (EFR) ist ein Substandard für die Codierung und Decodierung von Sprachsignalen in GSM-Netzen. Der EFR-Codec erhöht die *Sprachqualität*, so, dass diese in etwa der Sprachqualität des Festnetzes entspricht. Der EFR-Codec hat in etwa die gleiche Datenrate wie ein Full-Rate-Codec von GSM. Der *Sprachcodec* arbeitet mit *ACELP* und hat eine Bitrate von 12,2 kbit/s für *Sprache*. Das EFR-Verfahren hat gegenüber dem Full-Rate-Verfahren (FR) und dem Half-Rate-Verfahren (HR) eine hohe Sprachqualität.

Fernsprechen, Fe

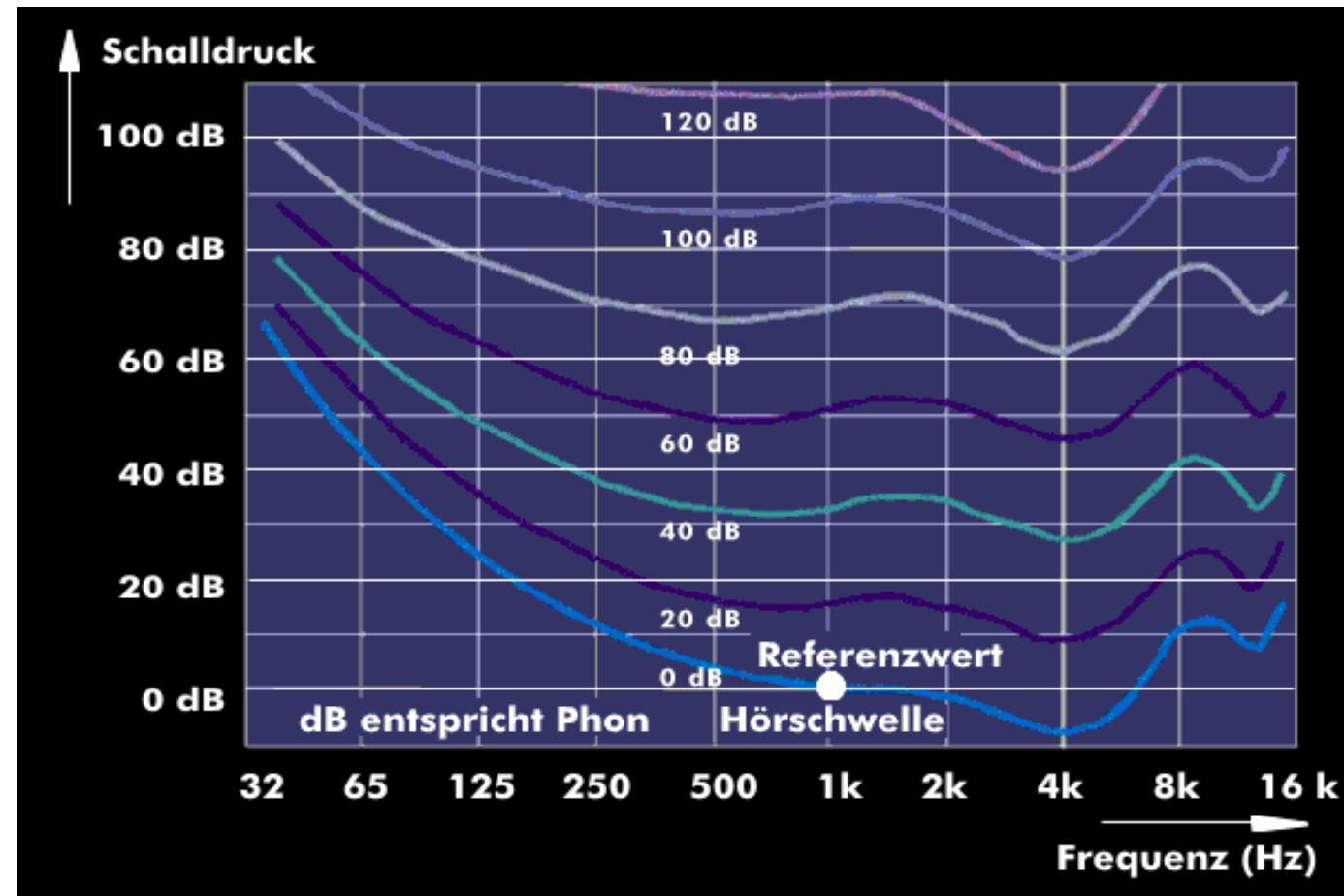
Fernsprechen ist *Sprachkommunikation* zwischen entfernten Teilnehmern über Telekommunikationsnetze. Das Fernsprechen oder die *Telefonie* erfolgt mit Telefonen, die die *Sprache* in elektrische Signale umwandeln, die dann über das Fernsprechnet übertragen werden.

Aufgrund der hohen Redundanz der menschlichen Sprache benutzt man aus Gründen der Frequenzökonomie ein möglichst geringes Frequenzband. Für die Bandbreite des Frequenzbandes waren die Spracherkennbarkeit der Teilnehmer und eine ausreichende Satz- und *Silbenverständlichkeit* ausschlaggebend. Diese liegt bei 3,1 kHz, und zwar im Frequenzbereich von 300 Hz bis 3.4 kHz. Unter Verwendung von Sprachspeichern kann Sprachkommunikation auch zeitversetzt erfolgen.

- G.107** Die ITU-T-Empfehlung G.107 beschreibt mit dem *E-Modell* ein Berechnungsmodell für die Planung und Bewertung der Übertragungsqualität von Übertragungsnetzen. Mit diesem Berechnungsmodell kann die *Sprachqualität*, die dem Nutzer bei einer Ende-zu-Ende-Verbindung zur Verfügung steht, ermittelt werden. Das Ergebnis ist eine objektive Bewertung der Übertragungsqualität unter Berücksichtigung aller, die Übertragungsqualität beeinflussender Faktoren.
- G.114** In der ITU-T-Empfehlung G.114 werden die *Verzögerungszeiten* und *Paketverluste* bei der paketvermittelten *Sprachkommunikation* spezifiziert und bewertet. Die Bewertung der *Sprachqualität* kann subjektiv mit dem *MOS-Wert* erfolgen oder objektiv mit dem aus Testverfahren ermittelten *R-Faktor*. Nach der G.114-Bewertung stuft man die Sprachqualität als inakzeptabel, noch akzeptabel und gut ein. G.114 hat den Titel "One-way Transmission Time" und gibt an, wie groß die Paketverluste und die Verzögerung für die G.114-Bewertung sein darf. Danach ist eine *Paketverlustrate* von bis zu 5 % nicht wahrnehmbar und die Verzögerungszeit sollte bei einer akzeptierbaren Sprachqualität unter 150 ms liegen. Bei größer 400 ms ist die Qualität inakzeptabel.

Hörcharakteristik *range of audibility*

Das menschliche Ohr hat ein frequenz- und altersabhängiges Hörempfinden: die so genannte Hörcharakteristik. Der hörbare Frequenzbereich ist lautstärkeabhängig und liegt bei Kindern im Frequenzbereich zwischen 20 Hz und 20 kHz. Die Hörcharakteristik ändert sich mit unterschiedlichen Schallpegeln und hat ihre höchste Sensitivität bei Frequenzen zwischen 2 kHz und 4 kHz. Bei diesen



Hörcharakteristik mit Hörschwelle

Frequenzen erreicht das Gehör einen maximalen Dynamikumfang bis zur Schmerzschwelle von etwa 130 dB. Töne mit höheren Frequenzen als 4 kHz werden ebenso wie tiefere Töne bei gleichem Schalldruck als leiser empfunden. Des Weiteren sind bestimmte Frequenzen oberhalb von 16 kHz und unterhalb von 30 Hz nicht hörbar. Die empfundene Lautstärke, die in Phon angegeben wird, ist abhängig von dem akustischen Signal und dessen Signalfrequenz.

Der Bezugswert für den Schalldruck ist 0 dB bei einer Frequenz von 1 kHz. Je nach Tonhöhe werden die Signale mit gleichem Schalldruck unterschiedlich laut empfunden. Dies wird durch die Linien, die den gleichen Schalldruck aufweisen, in der Hörcharakteristik verdeutlicht. Die geringste wahrnehmbare Lautstärke repräsentiert die *Hörschwelle*. Die Hörschwelle ist eine frequenzabhängige Kennlinie, die den niedrigsten wahrnehmbaren Schalldruck in Bezug zur Frequenz darstellt. Die höchste wahrnehmbare Lautstärke ist die Schmerzschwelle, die bei Schalldrücken von etwa 130 dB liegt. Der Bereich zwischen der Hörschwelle und der Schmerzschwelle wird Hörfeld genannt.

Hörschwelle
HL, hearing threshold

Die Hörschwelle (HL) charakterisiert die geringste Lautstärke, die ein Hörender wahrnehmen kann. Es handelt sich dabei um eine empirisch ermittelte Kennlinie der *Hörcharakteristik*, die den niedrigsten wahrnehmbaren Schalldruck über den hörbaren Frequenzbereich darstellt. Der Referenzwert beträgt bei 1.000 Hz 0 Dezibel (dB). Die Hörschwelle von 0 dB liegt bei einer Schalleistung (Pa) von 20 μ Pa, was 2×10^{-5} N/qm entspricht. Sie ist frequenzabhängig und steigt bei tiefen und hohen Tönen stark an. Das bedeutet, dass tiefe Töne, wenn sie mit gleicher Lautstärke empfunden werden sollen, einen wesentlich höheren Schalldruck benötigen. So muss ein tiefer Basston mit einem etwa 50 dB höheren Schalldruck abgestrahlt werden damit er genau so laut empfunden wird, wie ein 1-kHz-Ton. Die Hörschwelle ändert sich mit der Lautstärke. Bei lauten Tönen steigt die gesamte Kennlinie zu höheren Pegeln hin an, bei leisen Tönen sinkt sie ab und wird empfindlicher.

LD-CELP, low delay, code excited linear prediction

Low *Delay*, Code Excited Linear Prediction (LD-*CELP*) ist ein Algorithmus zur *Sprachkompression* in Leitungen mit einer Übertragungsrate von 16 kbit/s bzw. einer Kompression von 4:1, beschrieben im ITU-Standard G.728. Das Verfahren zeichnet sich durch geringe Signalverzögerungszeiten aus, die bei 625 µs liegen. Die durch den *MOS-Wert* ermittelte *Sprachqualität* ist als sehr gut zu bezeichnen und liegt bei 4,1.

MOS, mean opinion score *MOS-Wert*

Der Mean Opinion Score (MOS) ist ein subjektiver Bewertungsmaßstab für die Übertragung von *Sprache*. Er bietet die Möglichkeit, die Übertragungsqualität für unterschiedliche Sprachcodierungen miteinander zu vergleichen.

Der MOS-Wert ist ein dimensionsloser Wert zwischen eins und fünf, der für die *Sprachqualität* steht; wobei der Wert »eins« eine mangelhafte Sprachqualität repräsentiert, bei der keine Verständigung möglich ist, der Wert »fünf« hingegen für eine exzellente Übertragungsqualität steht, die nicht von dem Original zu unterscheiden ist.

Der MOS-Wert wird im Gegensatz zum objektiv ermittelten *R-Faktor* subjektiv ermittelt. Bei der Ermittlung des MOS-Werts spielt man Probanden Sprechproben vor, die diese bewerten. Die Bewertungen werden gewichtet und daraus werden die statistischen Ergebnisse ermittelt. In den ITU-Empfehlungen P.830 bis P.834 werden die Bewertungsmethoden verfeinert.

Wichtigste Qualitätskriterien für die Übermittlung von Sprachinformationen sind die *Verzögerungszeiten*, *Bitfehlerraten*, *Echos* und *Jitter*. Da das Ohr auf Klangschwankungen und Sprachunterbrechungen sensibel reagiert, sollten die Verzögerungszeiten annähernd konstant sein. Die Sprachqualität wird durch die Verzögerung während der Übertragung nicht beeinträchtigt, es verschlechtert sich lediglich die Gesprächsqualität. Bitfehler hingegen wirken sich durch Knackgeräusche aus.

Echos entstehen in analogen Systemen am Übergang von Vierdraht- auf Zweidrahttechnik und irritieren den Sprecher durch die Sprachreflexion, worunter die *Verständlichkeit* leidet.

Paketverlust *packet loss*

Bei der Datenpaketübertragung werden Datenpakete beschädigt, von überlasteten Routern nicht weitergeleitet oder finden den Empfänger nicht oder nach einer zu großen *Verzögerung*. Solche beschädigte und nicht benutzbare Datenpakete werden verworfen und bringen ihre Nutzdaten nicht zum Empfänger. Vor allem bei Echtzeitdiensten wie der Internettelefonie oder Videokonferenzen verschlechtern verzögerte Datenpakete oder der Paketverlust die Bild- und *Sprachqualität*. Ein nachträgliches Einfügen von verspätet eintreffenden Datenpaketen ist bei Echtzeitanwendungen nicht möglich. Der Verlust einzelner Datenpakete ist nicht unbedingt wahrnehmbar, allerdings verschlechtert der Paketverlust bei steigenden Verzögerungszeiten die Sprachqualität.

Die ITU-T hat in der Empfehlung *G.114* die *Paketverlustrate* für die Sprachqualität spezifiziert. Danach wird eine Paketverlustrate (PLR) bis zu 5 % kaum wahrgenommen.

Paketverlustrate *PLR, packet loss rate*

Die Paketverlustrate (PLR) gibt den prozentualen Anteil an verloren gegangenen Datenpaketen auf einer Übertragungstrecke wieder. Als Paketverlustrate dient dabei das Verhältnis aus Anzahl der *Paketverluste* zu der Gesamtzahl aller übertragenen Datenpakete.

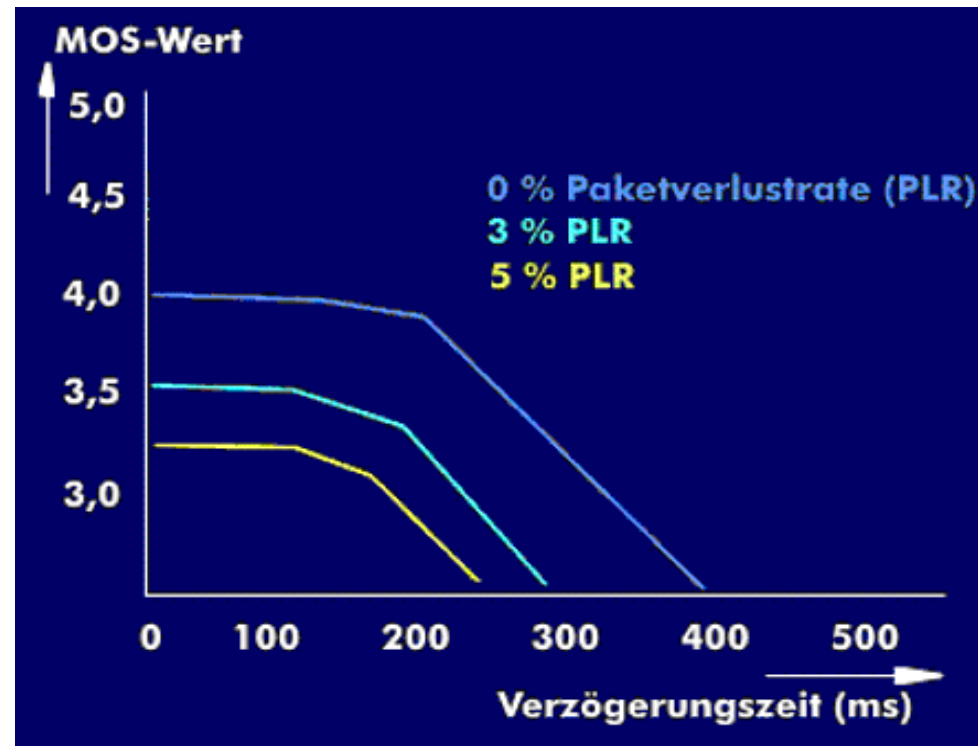
Die Paketverlustrate ist mitentscheidend für die *Sprachqualität* von paketvermittelter *Sprache*, wie VoIP,

MOS-Wert	Sprachqualität
5	excellent
4	gut
3	ordentlich
2	mäßig
1	mangelhaft

MOS-Werte mit entsprechender Sprachqualität

MOS-Wert in Abhängigkeit von der Paketverlustrate (PLR) beim G.729-Codec

PAMS, perceptual analysis measurement system



VoLAN oder VoWLAN. Sie ist von der ITU-T in der Empfehlung G.114 spezifiziert. Außerdem wird die Paketverlustrate in den QoS-Klassen der ITU-T angegeben. Bei hohem Verkehrsaufkommen, bei Überlast und bei Routingproblemen kann die Paketverlustrate im Internet auf bis zu 20 % ansteigen und die Übertragung massiv beeinträchtigen.

Perceptual Analysis Measurement System (PAMS) ist eine von mehreren Methoden zur Ermittlung der objektiven *Sprachqualität* in der *Telefonie*. PAMS funktioniert ähnlich wie das *Perceptual Speech Quality Measurement* (PSQM) und bietet eine objektive und reproduzierbare Methode für die

Qualitätsbewertung von Sprachübertragungen. Der Unterschied zwischen beiden Methoden liegt in dem psychoakustischen Modell und der Priorisierung der Klarheit der *Sprache* bei PAMS. Eine weitere Methode neben dem *Perceptual Speech Quality Measurement* (PSQM) ist die *Perceptual Evaluation of Speech Quality* (PESQ).

PESQ, perceptual evaluation of speech quality

Perceptual Evaluation of Speech Quality (PESQ) ist neben *Perceptual Analysis Measurement System* (PAMS) und *Perceptual Speech Quality Measurement* (PSQM) eine weitere Methode zur objektiven Bewertung der *Sprachqualität* in der *Telefonie*. PESQ ist in der ITU-Empfehlung Q.862 beschrieben und basiert auf den realen Bedingungen für eine Ende-zu-Ende-*Sprachkommunikation*. Das Verfahren wird für VoIP eingesetzt und berücksichtigt u.a. *Paketverluste*, *Rauschen*, die Codierung. Die Ergebnisse von PESQ werden in einer Skala von 4,5 bis -0,5 wiedergegeben, wobei der höchste Wert die beste Sprachqualität repräsentiert.

<http://www.pesq.org>

POTS, plain old telephone service
Telefonie

Telefonie ist eine *Sprachkommunikation* über öffentliche Netzwerke. Die *Sprache* wird dabei in elektrische Signale umgewandelt, analog oder digital übertragen und empfangsseitig in Hörsignale rückgewandelt. Mit POTS (Plain Old Telephone Service) bezeichnet man die klassische analoge Telefonie, die eine Bandbreite von 3,1 kHz hat und im Frequenzbereich von 300 Hz bis 3,4 kHz liegt. Dieser Frequenzumfang ist so gewählt, dass die Sprache verständlich ist und der Sprechende mit seinen Stimmcharakteristiken erkannt werden kann, mehr aber auch nicht. Neben der *Silbenverständlichkeit*, die bei über 90 % liegt, ist die Satzverständlichkeit mit etwa 99 % entscheidend. Da bei der Telefonie die Frequenzbereiche unterhalb von 300 Hz und oberhalb von 3,4 kHz nicht übertragen

werden, leiden *Sprachqualität* und Sprachverständlichkeit. Dies drückt sich bei den fehlenden tiefen Frequenzen dadurch aus, dass das Sprachvolumen und die -natürlichkeit leiden. Bei den fehlenden hohen Frequenzen ist eine Einschränkung der Silbenverständlichkeit auszumachen. Um diese Nachteile zu beheben wurden Verfahren entwickelt mit denen die Sprachbandbreite künstlich auf einen Frequenzbereich zwischen 50 Hz und 7 kHz erweitert wurde. Diese Verfahren werden noch nicht eingesetzt.

PSQM, perceptual speech quality measurement

Perceptual *Speech* Quality Measurement (PSQM) ist eine von mehreren Methoden zur Ermittlung der objektiven *Sprachqualität* für die *Telefonie*. PSQM ist ein von der ITU-T unter dem Standard P.861 beschriebenes Verfahren, mit dem niederbitratige *Sprachcodecs* getestet werden. Dabei wird ein spezifiziertes Sprachsignal an einem Ende einer Verbindung eingespeist und am anderen Ende wird eine objektive Bewertung des Ausgangssignals mit speziellen Berechnungsformeln vorgenommen. Man nutzt dabei ein psychoakustisches Modell um das subjektive Empfinden mathematisch nachzubilden. Das Ergebnis ist ein objektiver Qualitätswert, der so genannte PSQM-Wert, der zwischen 0 und 6,5 liegen kann und in den *MOS-Wert* übertragen wird. Ein niedriger PSQM-Wert entspricht einer hohen Sprachqualität, ein hoher PSQM-Wert einer minderen bis schlechten Sprachqualität.

Weitere Verfahren zur Ermittlung der Sprachqualität sind das *Perceptual Analysis Measurement System* (PAMS) und die *Perceptual Evaluation of Speech Quality* (PESQ). Die beiden Verfahren PSQM und PAMS arbeiten nach mathematischen Algorithmen und basieren auf der Auswertung natürlicher männlicher und weiblicher Sprachmuster.

R-Faktor *R, transmission quality rating*

Der R-Faktor ist ein objektiv ermittelter Wert für die übertragene *Sprachqualität* in Fernsprechnetzen. Er wird zur Quantifizierung der Sprachqualität benutzt und bestimmt die Güte der Sprachübertragung. Beschrieben ist diese Standardmethode im *E-Modell* der ITU-T, das in G.107 spezifiziert ist. Neben dem objektiven R-Wert

R-Faktor	Bewertung der Sprachqualität
100	Exzellent, gut verständlich
G.107 → 94 Default-Wert	
80	Gut, Sprache kann verstanden werden
60	Ordentlich, beim Zuhören ist eine gewisse Konzentration erforderlich
50	Mäßig, man kann die Sprache nur mit hoher Konzentration verstehen
50 bis 0	Mangelhaft, es ist keine Verständigung möglich

gibt es noch den subjektiv ermittelten *MOS-Wert*, die beide in einer gewissen Korrelation zueinander stehen. Der R-Faktor kann Werte zwischen 0 und 120 annehmen, wobei der R-Wert von 100 und darüber einem MOS-Wert von 5 entspricht. Bedingt durch die Sprachkonvertierung in ein digitales Signal und dessen Rückwandlung liegt das theoretische Maximum des R-Faktors auf 93,2 begrenzen. Das entspricht einem MOS-Wert von 4,41. Der in G.107 definierte

Objektive Sprachbewertung durch den R-Faktor

Default-Wert für den R-Faktor liegt bei 94. In der Praxis werden aber bedingt durch die Codecs maximale R-Werte von etwas über 80 erreicht.

Neben dem benutzten Codec beeinflussen verschiedene übertragungstechnische Parameter den R-Faktor. So die *Verzögerungszeiten*, das *Signal-Rausch-Verhältnis*, der *Jitter*, *Echos* und *Paketverluste*.

RASTI, rapid speech transmission index

Für die Bewertung der *Sprachverständlichkeit* in beschallten Räumen gibt es mehrere Messverfahren.

Neben dem Rapid Speech Transmission Index (RASTI) gibt es den Speech Transmission Index (STI), den Speech Intelligibility Index (SII) oder *ALcons*.

RASTI wurde als IEC-Standard 60268-16 standardisiert und bewertet die Sprachverständlichkeit anhand von moduliertem *Rauschen*. Es ist eine einfachere Methode zu dem relativ aufwendigen Speech Transmission Index (STI).

RASTI arbeitet mit einem modulierten Rauschsignal, das über die Lautsprecher abgestrahlt und von Messmikrofonen aufgenommen und analysiert wird. Die von den Mikrofonen empfangenen Signale werden mit den Lautsprechersignalen verglichen und frequenzmäßig gewichtet. Im Gegensatz zu dem STI-Index erfolgt die frequenzmäßige Messung nur in zwei Oktaven mit den Mittenfrequenzen von 500 Hz und 2 kHz. Zur Modulation benutzt RASTI ein Signal, das einem Sprachsignal ähnlich ist und trifft anhand des veränderten Modulationsindex die Aussage über die *Sprachverständlichkeit*.

Rauschen *N, noise*

Rauschen sind statistisch verteilte Schwankungen einer physikalischen Größe, die durch stochastische Prozesse entstehen. In der Elektronik werden diese stochastischen Prozesse von dem Stromfluss in aktiven und passiven Bauteilen generiert. Es handelt sich dabei um unkoordinierte temperaturbedingte Elektronenbewegungen.

Rauschen wird im Wesentlichen durch die Verteilung der Rauschenergie über die Bandbreite bestimmt. Ist die Rauschenergie konstant über ein Frequenzband, spricht man von weißem Rauschen, ist keine Rauschenergie vorhanden, handelt es sich um schwarzes Rauschen. Rauschen, dessen spektrale Leistungsdichte von der des weißen Rauschens abweicht, nennt man farbiges Rauschen und ordnet diesem, je nach Frequenzabhängigkeit, bestimmte Farben zu. Das farbiges Rauschen, das nicht exakt spezifiziert ist, ist dadurch gekennzeichnet, dass die Rauschenergie eine über- oder unterproportionale Frequenzabhängigkeit aufweist. Diesen Rauscharten sind Farben zugeordnet: rot/braunes Rauschen, blaues Rauschen, violetttes Rauschen und rosa Rauschen. Des Weiteren kennt man das orange Rauschen, das grüne Rauschen oder das graue Rauschen.

Im Federal Standard 1037C Telecommunications werden das weiße Rauschen, das schwarze Rauschen, das rosa Rauschen und das blaue Rauschen definiert.

Da Rauschen abhängig ist von der Bandbreite, beeinflusst es das Nutzsignal. Das Verhältnis von Nutzsignal zu Rauschsignal nennt man *Signal-Rausch-Verhältnis* (SNR).

Ein Maß für die in Satelliten-Empfangsanlagen gemessene Rauschtemperatur ist das Kelvin.

SBR, spectral band replication

Audiocodecs nutzen neben verschiedenen Kompressions-Algorithmen auch diverse andere Techniken um bei niedrigen Bitraten eine möglichst hohe Klangqualität zu erzielen und dabei die Dateigrößen möglichst

klein zu halten.

Das SBR-Verfahren (Spectral Band Replication), das in aacPlus und mp3pro angewendet wird, ist eine Codierungs-Technologie, die sich durch eine hohe Klang- und *Sprachqualität* bei niedrigen Bitraten auszeichnet.

Das SBR-Verfahren setzt bei Tönen von über 5 kHz ein, die aus der normalen Decodierung gewonnen und mittels Spectral Band Replication rekonstruiert werden. Durch diese Rekonstruktion brauchen die hohen Töne, die eine größere Datenmenge repräsentieren, nicht mit übertragen werden. Die Ton-Rekonstruktion basiert auf den tieferen Frequenzen, die von dem normalen Decoder erzeugt werden. Damit eine einwandfreie Rekonstruktion der Audiosignale stattfinden kann, werden zusätzlich zu den übertragenen Audiodaten noch Steuerdaten im Bitstrom übertragen. Die Rekonstruktion arbeitet sehr effizient mit den Harmonischen und ermöglicht die richtige Rekonstruktion der Signalform in Bezug auf dessen Zeit- und Frequenzverhalten.

Signal-Rausch-Verhältnis, S/N *SNR, signal to noise ratio*

Das Signal-Rausch-Verhältnis (SNR) ist der Quotient aus der Leistung des übertragenen Nutzsignals zur Leistung des Rauschsignals und ein Maß für die Reinheit eines Signals. Da das Verhältnis zwischen Nutzsignal und Rauschsignal mehrere Zehnerpotenzen umfassen kann, wird das Signal-Rausch-Verhältnis im logarithmischen Maßstab angegeben und dafür wird das Dezibel (dB) benutzt.

Berechnung des Signal-Rausch-Verhältnisses

$$\text{SNR (dB)} = 10 \log \frac{\text{Leistung Nutzsignal}}{\text{Leistung Rauschsignal}}$$

Das Signal-Rausch-Verhältnis ist ein wichtiger Kennwert für die Dynamik von Vierpolen; so von Verstärkern, A/D-Wandlern und Mikrofonen. Die Dynamik kann immer nur so groß sein, wie das SNR-Verhältnis, da ja sonst der Verstärker bereits das

Rauschen verstärken würde.

Das Signal-Rausch-Verhältnis kann durch bestimmte Maßnahmen verbessert werden. Neben der Erhöhung des Nutzsignals werden Expandertechniken eingesetzt, bei denen Nutzsignale mit geringem Pegel vor der Übertragung mit höherem Pegel übertragen und nachher wieder dekomprimiert werden. Auch bietet sich der Einsatz von Filtern an, die das Rauschsignal ab einer bestimmten Frequenz begrenzen, ohne das Nutzsignal zu beeinträchtigen.

SII, speech intelligibility index *SII-Index*

Für die Bewertung der *Sprachverständlichkeit* gibt es mehrere Bewertungsverfahren. So den *Speech Transmission Index (STI)*, den relativ betagten *Articulation Index (AI)* und den aus dem STI-Index abgeleiteten *Speech Intelligibility Index (SII)*.

Die meisten Bewertungsverfahren basieren auf dem Vergleich der vom Lautsprecher abgestrahlten Signale mit denen, die das Mikrofon aufnimmt. Als Lautsprechersignale werden bei vielen Verfahren, so auch beim SII-Index, modulierte Audiosignale generiert. Die vom Mikrofon aufgenommenen Schallwellen werden über bestimmte Filtercharakteristiken gefiltert und mit den Lautsprechersignalen verglichen. Aus den Abweichungen in der Modulationstiefe und dem Anteil an Stör- und Hintergrundgeräuschen wird dann die Bewertung für mehrere Frequenzbereiche vorgenommen und die Ergebnisse werden nach einem Algorithmus gemittelt.

Bewertung der Sprachverständlichkeit nach dem STI-, SII-Index und ALcons

STI-Index SII-Index	Sprachverständlichkeit	Alcons
0 bis 0,3	Nicht akzeptierbar, unverständlich	100 % bis 33 %
0,3 bis 0,45	Schlecht	33 % bis 15 %
0,45 bis 0,6	Genügend	15 % bis 7 %
0,6 bis 0,75	Gut	7 % bis 3 %
0,75 bis 1,0	Ausgezeichnet	3 % bis 0 %

Der SII-Index ist aus dem STI-Index abgeleitet. Es handelt sich dabei um ein maschinelles Bewertungsverfahren für die Sprachverständlichkeit, das von ANSI standardisiert wurde. Nach dem ANSI-Standard sind vier Mess-Prozeduren erlaubt, von denen jede für sich eine unterschiedlichen Anzahl und Breite an Frequenzbändern vorsieht. Diese Frequenzbänder werden aus 21 kritischen Frequenzbändern, 18 Terz-Bändern, 17 gleich verteilten kritischen Bändern und 6 Oktav-Bändern ausgewählt.

Wie der STI-Index kann auch der SII-Index

Werte zwischen "0" und "1" annehmen. "0" steht dabei für nicht akzeptable Sprachverständlichkeit, "1" für exzellente.

Der SII-Index zeigt als maschinelles Verfahren eine gute Übereinstimmung mit statischen Bewertungen, vor allem unterstützt er breitbandige Bewertungen zwischen 150 Hz und 8,5 kHz, aber auch schmalbandige Bewertungen mit den kritischen Bändern.

Sprachcodec voice codec

Sprachcodecs sind software- oder hardwaremäßige Funktionseinheiten in denen *Sprache* digitalisiert und komprimiert wird, damit sie über digitale und paketvermittelte IP-Netze übertragen werden kann. In Sprachcodecs werden verschiedene Verfahren der *Sprachkompression* angewandt, die auf die jeweiligen Anforderungen an die Netze abgestimmt sind. Ziel aller Kompressionsverfahren ist es, die Sprache mit möglichst niedriger Übertragungsrates in hoher Qualität übertragen zu können. Da bei der Quantisierung der Sprache Verzögerungen auftreten, die die *Sprachqualität* beeinträchtigen, ist bei der Auswahl der Sprachcodecs immer mit Kompromissen zu arbeiten. Die ITU hat in ihren G-Empfehlungen G.721 bis G.729 verschiedene Sprachcodecs mit unterschiedlichen Kompressionsalgorithmen standardisiert. Die Sprachqualität von Sprachcodecs wird wesentlich durch die bei der Quantisierung auftretenden *Verzögerungszeiten* und den *Jitter* beeinträchtigt, aber ebenso durch die Dekompression. Mit höher werdender Sprachkompression, verringert sich die Sprachqualität. An bekannten Verfahren für die Sprachkompression sind zu nennen die Pulscodemodulation (PCM), ADPCM, CELP, ACELP, LD-CELP, MPMLQ, HVXC u.a.

Sprache speech

Sprache ist eine Kommunikationsart mit der sich Menschen untereinander verständigen. Sie besteht aus einzelnen Lauten, die aneinander gereiht Wörter ergeben. Der Frequenzbereich für die menschliche Sprache liegt zwischen 300 Hz und etwa 5 kHz. Die Laute sind zwischen 10 ms und 50 ms lang und bestehen aus Tönen im angegebenen Frequenzbereich. Charakteristische Ausprägungen im Laut- und Frequenzbereich erzeugen die unterschiedlichen und wiedererkennbaren Sprachmerkmale.

Die Sprache umfasst nur einen kleinen Bereich der *Hörcharakteristik*. Dieser ist durch den Frequenzbereich zwischen 300 Hz und 5 kHz und einer Dynamik von etwa 30 dB gekennzeichnet.

Sprachfrequenz VF, voice frequency

Unter der Sprachfrequenz (VF) versteht man den Frequenzbereich, der das menschliche Sprechen umfasst. Dieser Frequenzbereich liegt zwischen 300 Hz und 4 kHz. Der Frequenzbereich in dem es um *Verständlichkeit* und das Erkennen der Stimmcharakteristiken geht, liegt zwischen 300 Hz und 3,4 kHz. Dieser Frequenzbereich wird für die *Sprachkommunikation* in der *Telefonie* benutzt.

Sprachkommunikation voice communications

Die traditionelle Sprachkommunikation zwischen entfernten Teilnehmern ist über das *Fernsprechen* seit Jahren realisiert. Diese gegenseitige akustische Verständigung von Mensch zu Mensch über größere Entfernungen, ist die meistbenutzte Telekommunikationsform. Für diese Kommunikationsform gibt es das leitungsvermittelte Fernsprechnet, über das Sprachverbindungen zwischen zwei und mehreren Partnern hergestellt werden können.

Die Sprachkommunikation erfordert eine möglichst verzögerungsfreie Übertragung, damit der kontinuierliche Datenstrom für die *Sprache* beim Empfänger originalgetreu eintrifft. So ist die *Verständlichkeit* der übertragenen Sprache auch dann noch gegeben, wenn die Übertragung durch Aussetzer, *Jitter*, *Verzögerungen* und Störeinflüsse wie Knistern oder *Rauschen* beeinträchtigt wird.

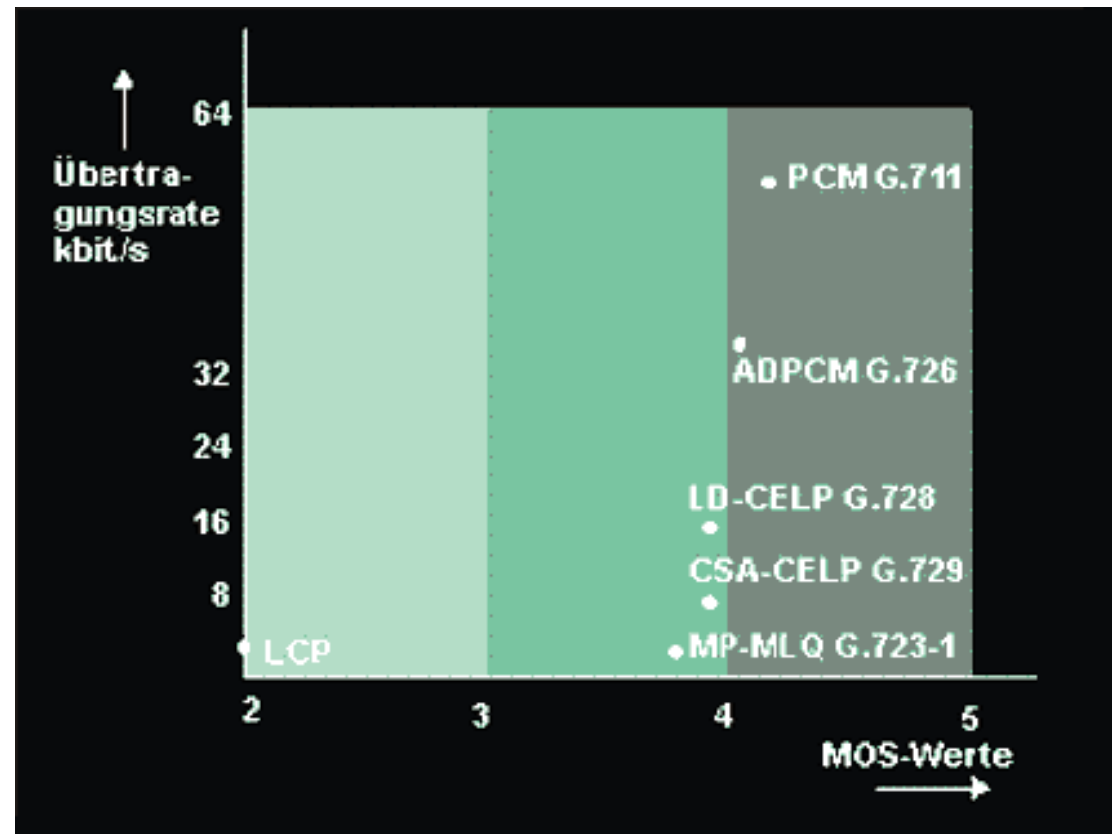
Die Bewertung der *Sprachqualität* kann objektiv und subjektiv erfolgen. Eine bekannte subjektive Bewertung ist der Mean Opinion Score (MOS). In Verbindung mit dem von der ITU unter G.107 standardisierten *E-Modell* kann dieser auch in Simulationen bestimmt werden.



ITU-Empfehlung G.114 für Verzögerungszeiten bei Sprachübertragungen

Wichtigste Qualitätskriterien für die Übermittlung von Sprachinformationen sind die *Verzögerungszeiten*, *Bitfehlerraten*, *Echos* und *Jitter*. Da das Ohr auf Klangschwankungen und Sprachunterbrechungen sensibel reagiert, sollten die Verzögerungszeiten annähernd konstant sein. Die Sprachqualität wird durch die Verzögerung während der Übertragung nicht beeinträchtigt, es verschlechtert sich lediglich die Gesprächsqualität. Bitfehler hingegen wirken sich durch Knackgeräusche aus. Eine Bitfehlerrate von 10×10^{-2} macht sich ein stark störendes Prasseln bemerkbar, bei 10×10^{-3} löst sich das Prasseln in einer dichten Folge von Knackgeräuschen auf, die bei einer Bitfehlerrate von 10×10^{-4} in einzelne Knackgeräusche übergeht. Bei 10×10^{-5} sind nur noch einzelne Knackgeräusche hörbar und Bitfehlerraten von 10×10^{-6} beeinträchtigen die Sprachübertragung überhaupt nicht.

Sprachkompression *speech compression*



Sprachqualität in Abhängigkeit vom Kompressionsverfahren

Um echtzeitorientierte Sprachanwendungen über Datenpaketnetze übertragen zu können, müssen die zu übertragenden Daten komprimiert werden. Für die Sprachkompression hat die ITU Standards für Codierverfahren verabschiedet, die abhängig von der nutzbaren Bandbreite unterschiedliche Qualität in der *Sprachverständlichkeit* bieten. Diese Kompressionsverfahren sind Bestandteil der Protokollfamilie H.323, zu der mehrere *Sprachcodecs* zählen.

Als Beurteilungskriterium für die Qualität der codierten *Sprache* hat die ITU eine Messgröße definiert, die aus der durchschnittlichen Bewertung verschiedener Sprachmuster durch

mehrere Personengruppen errechnet wird: Diese Bewertungsgröße heißt »Mean Opinion Score« (*MOS*). Die *MOS*-Skala reicht von 0 bis 5, wobei der obere Wert eine gute *Sprachverständlichkeit* repräsentiert. In Abhängigkeit von dem verwendeten Codierverfahren kann die effektive, für die Sprachübertragung genutzte Bandbreite um bis zu 90% reduziert werden. Dabei ist zu berücksichtigen, dass alle Verfahren einen relativ umfassenden Overhead für die Übertragung benötigen. Dieser Overhead kann den Datenanteil für die Sprache um ein Mehrfaches übertreffen.

Bei den ITU-T-Standards handelt es sich um die G-Empfehlungen G.711 bis G.729, die mit unterschiedlichsten Codierverfahren, Bandbreiten, *Verzögerungszeiten* und *MOS*-Werten arbeiten.

An Verfahren für die Sprachkompression sind zu nennen: ADPCM, *CELP*, GSM 06.10, MPMLQ, HVXC und die LPC-Codierung.

Sprachqualität *voice quality*

Die Qualität der Sprachübertragung in leitungsvermittelten Fernsprechnetzen und paketvermittelten IP-Netzen bestimmt im Wesentlichen die Kommunikation zwischen den Gesprächspartnern.

Die Sprachqualität wird maßgeblich von den *Sprachcodecs* bestimmt und durch Codierungen und *Sprachkompression*, durch *Echos*, *Paketverluste*, Störeinflüsse und Verzögerungen beeinträchtigt. Codecs mit hoher Kompression benötigen weniger Bandbreite, tendieren aber zu höheren Fehlerraten und Qualitätsverlusten. Unabhängig von der Art der Beeinträchtigung gibt es objektive und subjektive Bewertungen für die Sprachqualität. Die ITU-T hat für die objektive Bewertung mit dem *E-Modell* ein Verfahren entwickelt, das im Standard G.107 spezifiziert ist und das auf reproduzierbaren Störgrößen basiert. Für die Quantifizierung der Sprachqualität und die Güte der Sprachübermittlung gibt es den *R-Faktor*. Die

subjektive Bewertung drückt sich in den so genannten *MOS-Werten* aus und in den von der ITU in den Empfehlungen Q.800 und Q.830 verfeinerten Methoden zur Bewertung der Sprachqualität. Neben dem MOS-Wert spielt der *Sprachkommunikation* mittels VoWLAN der R-Faktor eine entscheidende Rolle. Dieser hängt von der *Paketverlustrate*, dem *Jitter* und der *Verzögerung* ab. Zu den objektiven Bewertungsmethoden für die Sprachqualität gehören Messmethoden für den Störspannungsabstand, die Echomessung, das *Perceptual Speech Quality Measurement (PSQM)*, das *Perceptual Analysis Measurement System (PAMS)* und die *Perceptual Evaluation of Speech Quality (PESQ)*.

Sprachverständlichkeit *speech intelligibility*

Von *Sprachverständlichkeit* spricht man bei der Kommunikation und auch bei der Beschallung von Räumen. In der Kommunikation ist die Verständlichkeit durch Codierungsverfahren, Bandbreiten, Abtastraten, Auflösungen, *Verzögerungszeiten*, *Bitfehlerraten*, *Echos* und *Jitter* geprägt, wohingegen in der Akustik die *Sprachqualität* in beschallten Räumen durch die Raumform, Raumgröße, durch Reflexionen, Absorptionen, Hintergrundgeräusche, Abstrahlcharakteristiken der Lautsprecher, Position der *Absorber* und des Zuhörers beeinflusst wird.

In der *Sprachkommunikation* gibt es als bekanntestes Bewertungsverfahren für die *Telefonie* den Mean Opinion Score (*MOS*), der sich aus der mittleren Bewertung verschiedener Sprachmuster durch mehrere Personen errechnet.

Was die Beschallung von Räumen betrifft, so ist die Nachhallzeit ein wesentlicher Faktor, der die Hörsamkeit entscheidend beeinträchtigt. Je kürzer die Nachhallzeit desto besser die Sprachverständlichkeit in einem Raum. Bei langer Nachhallzeit kann der Nachhall die folgende Silbe beeinträchtigen und damit die Sprachverständlichkeit verschlechtern.

Für die Bewertung von beschallten Räumen gibt es mehrere maschinelle und statische Verfahren. Dazu gehören der *Speech Transmission Index (STI)*, *Rapid Speech Transmission Index (RASTI)*, *Speech Intelligibility Index (SII)*, der betagte *Articulation Index (AI)* und *ALcons*.

Sprachübertragungsindex *STI, speech transmission index*

In beschallten Räumen ist die *Sprachverständlichkeit* von diversen Faktoren abhängig, so von dem Nachhall, der Raumgröße und dessen Aufbau, der Richtcharakteristik der Lautsprecher, von Störgeräuschen und den Reflexionen im Raum. Für die Bewertung der *Sprachverständlichkeit* in beschallten Räumen gibt es mehrere Verfahren. *Speech Transmission Index (STI)* ist eines dieser Bewertungsverfahren, *Speech Intelligibility Index (SII)*, *Rapid Speech Transmission Index (RASTI)* und *ALcons* weitere.

Die STI-Messung ist ein maschinelles Verfahren und basiert auf der Sprachmodellierung mit einem Testsignal, dessen Charakteristik Sprachsignalen entspricht. Da Sprache als eine Modulationstechnik angesehen werden kann, wird das STI-Testsignal aus einer Grundwelle gebildet, die mit niedrigen Frequenzen moduliert wird. Die gesamte STI-Messung besteht aus vielen Modulationen, die den gesamten Sprachbereich abdecken und in der Modulationstiefe verändert werden. Gemessen wird die Modulationstiefe des empfangenen Tonsignals, das mit dem abgestrahlten Tonsignal verglichen wird. Aus dem Quotienten wird der dimensionslose STI-Index gebildet, der Werte zwischen "0" und "1" annehmen kann. Der STI-Wert "0" steht für ungenügende, der STI-Wert von "1" für exzellente Sprachverständlichkeit.

- Verständlichkeit**
audibility Die sprachliche Kommunikation zwischen zwei und mehreren Kommunikationspartnern, drückt sich, ebenso wie die audiophone Sprachübertragung über Rundfunk oder in beschallten Räumen in der Verständlichkeit aus.
1. Die Verständlichkeit ist ein Kennwert der Audio- und Telefontechnik in den die Deutlichkeit und die Silbenverständlichkeit eingehen. Die Audiotechnik kennt dafür noch die Begriffe Durchsichtigkeit und Klarheit. Die Interpretation der Verständlichkeit ist dienstspezifisch. So ist beim Telefonieren die Verständlichkeit bereits erreicht, wenn die Übertragungstechnik eine ausreichende Sprach- und Silbenverständlichkeit zur Verfügung stellt, die ja durch *Verzögerungszeiten*, *Bitfehlerraten*, *Echos* und *Jitter* hinreichend beeinträchtigt wird. Mit dem *MOS-Wert* hat die telefonische *Sprachkommunikation* ein Instrument für die Bewertung der *Sprachqualität*.
 2. Im Gegensatz zu der Verständlichkeit in der Kommunikationstechnik wird die Verständlichkeit in beschallten Räumen durch Reflexionen, Absorptionen und Nachhall geprägt. Diese frequenzabhängigen Faktoren beeinträchtigen die *Sprachverständlichkeit*, die maßgeblich durch Reflexionen bis 50 ms verbessert wird. Reflexionen, die unter 50 ms nach dem Direktsignal eintreffen, tragen zur Deutlichkeit bei. Für eine qualitative Bewertung der Sprachverständlichkeit gibt es mehrere Verfahren: den *Speech Transmission Index (STI)*, den *Rapid Speech Transmission Index (RASTI)*, den *Articulation Index (AI)*, den *Speech Intelligibility Index (SII)* und *ALcons*.
- Verzögerung**
DEL, delay Unter Verzögerungszeit (DEL) ist in der Kommunikation die Zeitspanne zu verstehen, um die ein Signal verzögert beim Empfänger eintrifft. Dieser Parameter ist von besonderer Bedeutung bei Echtzeit- und Multimedia-Anwendungen.
- ITU-Studien haben ergeben, dass die *Sprachqualität* sehr stark durch die Verzögerungszeiten und das *Echo* beeinträchtigt wird. Für das klassische Telefonnetz wurden die Ende-zu-Ende-Laufzeiten im nationalen Bereich auf 25 ms festgelegt; im internationalen Bereich auf 100 ms. Für das Echosignal gibt es zwei Grenzwerte von 30 ms und 150 ms. In diesem Bereich kann das störende Echo kompensiert werden. In paketvermittelten Netzen, so bei VoIP, kommt es bei der *Sprachkommunikation* generell zu vermittlungs- und netzspezifischen Verzögerungen. Im Audiocodec und der Paketierung im Sender kann die Verzögerung zwischen 20 ms und 25 ms betragen, für die Zwischenspeicherung im Netzbereich können Verzögerungen bis zu 700 ms auftreten und bei der Synchronisierung, Depaketierung und Decodierung nochmals bis zu 150 ms. Bei der Übertragung von grafischen Informationen in Echtzeit liegt der akzeptable Wert bei 30 ms. In Weitverkehrsnetzen sollte die Verzögerungszeit 100 ms nicht überschreiten, im LAN sollte sie unter 30 ms liegen. Die Verzögerungszeiten für Sprachübertragungen sind von der ITU in der Empfehlung *G.114* festgelegt.
- VG, voice grade** Voice Grade (VG) ist eine Klassifikation für eine Kommunikationsleitung, die im normalen Telefon-Dienst benutzt wird und über die Sprachsignale ohne merkbare Beeinträchtigung übertragen werden können. Die Klassifikation ist unabhängig von der Leitungslänge. Der übertragene Frequenzbereich liegt zwischen 300 Hz und 3,4 kHz. Dieses Frequenzband wurde so gewählt, dass die *Verständlichkeit* und die Erkennbarkeit der Teilnehmer gewährleistet sind.