Bild: Agilent

Sprachqualität objektiv testen

Von Paul Denisowski und Leo De Rosa

Die Sprachverständlichkeit ist maßgeblich dafür verantwortlich, ob sich Paket-basierte Sprachnetze durchsetzen werden. Dreh- und Angelpunkt für die Einführung dieser neuen Technik ist die Fähigkeit, die Sprachqualität erkennen und messen zu können.

Aus verschiedenen Gründen ist es wichtig, Sprachqualität objektiv quantifizieren zu können. Zunächst lässt sich damit die Sprachqualität eines Paket-basierten Netzes mit der des herkömmlichen Telefonnetzes vergleichen, denn letzteres gilt gewissermaßen als Messlatte für eine annehmbare Sprachqualität. Zum Zweiten ist es wichtig zu wissen, welche Auswirkungen auf die Sprachqualität Änderungen der Designs oder unterschiedliche Netzbedingungen verursachen, beispielsweise Jitter. Nicht zuletzt sind objektive Messwerte aus kommerzieller Sicht wertvoll. Die Daten ermöglichen den Netzbetreibern den Vergleich des eigenes Angebots mit dem der Wettbewerber; sie dienen auch als Grundlage für Vereinbarungen über die Dienstequalität von Sprachdiensten, so genannten Service Level Agreements (SLA).

Laufzeiten bestimmen die Sprachqualität

Wie ein Individuum Sprachqualität empfindet, hängt von drei Hauptfaktoren ab: Laufzeit, Echo und Verständlichkeit. Damit ein empfangenes Sprachsignal als annehmbar eingestuft wird, müssen alle drei Parameter bestimmte Grenzen einhalten.

Die Laufzeit, auch als Latenz bezeichnet, ist von den drei genannten Faktoren am leichtesten nachzuvollziehen und zu quantifizieren; es ist die Zeit, die Signale vom Sprecher zum Hörer brauchen. Für erdgebundene Anrufe variiert die Laufzeit im üblichen Telefonnetz kaum; sie bewegt sich in der Größenordnung von einigen Millisekunden und ist in erster Näherung von der Wegstrecke abhängig. Unter bestimmten Bedingungen kann die Laufzeit jedoch leicht Hunderte von Millisekunden betragen, sogar in einem herkömmlichen Telefonnetz. Am ehesten geschieht das bei einem Überseegespräch, das über einen Satelliten läuft. Hoch zum Satelliten und wieder zurück sind es etwas über 70.000 km, dafür brauchen Signale bei Lichtgeschwindigket fast 250 ms. Eine Verzögerung von mehr als 150 ms beeinträchtigt bereits den Fluss der Unterhaltung.

Die Teilnehmer unterbrechen einander oder sprechen durcheinander, bis sie merken, dass die andere Seite spricht. Bei genügend großer Verzögerung verkommt die Unterhaltung zum Halb-Duplex-Gespräch, bei der die Unterhaltung jeweils wechselseitig ausschließlich nur in eine Richtung läuft. Paket-basierte Sprachnetze (VoIP) erreichen typischerweise diese Schwelle von 150 ms Laufzeit oder liegen gar darüber. Die Verzögerungen haben verschiedene Ursachen; die Laufzeit in den Leitungen und Schaltstationen spielen eine Rolle, und die Zeit zur Codierung der Sprache. Zusätzlich ändert sich die Laufzeit in Paket-basierten Sprachnetzen gemeinhin mit Tageszeit, Wochentag oder Netzgegebenheiten.

Der Effekt des Echos bedeutet einfach, seine eigene Stimme zu hören. Echo tritt auf, wenn ein Teil des Sendesignals auf dem Empfangskanal erscheint. Eine erwünschte Form des Echos ist das lokale Echo, wenn die eigene Stimme ohne Verzögerung im Telefonhörer zu hören ist. Viele Anwender verwirrt diees; sie glauben dann, der Gesprächspartner könne sie auch nicht hören.

Begriffe und ihre Bedeutung

Bark-Skala: Eine nichtlineare Skala von Frequenzbereichen, die den ersten 24 kritischen Hörbändern entsprechen. Die Mittenfrequenzen der Bereiche unterscheiden sich am unteren Rand der Skala um lediglich 100 Hz (50 Hz, 150 Hz, 250 Hz ...). Am oberen Ende vergrößert sich der Frequenzabstand (4000 Hz, 4800 Hz, 5800 Hz, 7000 Hz, 8500 Hz ...).
Circuit switching: Leitungsvermittlung ist die herkömmliche Vermittlungstechnik im Telefonnetz. Eine Verbindung muss zwischen den Teilnehmern aufgebaut werden, bevor das eigentliche Gespräch beginnen kann. Wenn es endet, wird die Verbindung abgebaut und die belegte Ressource frei für andere Teilnehmer.
Codec: Ein Gerät, das analoge Sprachsignale digitalisieren kann und umgekehrt. Ein Codec unterscheidet sich von einem Analog-Digital- beziehungsweise Digital-Analog-Wandler dadurch, dass er Sprachsignale nicht nur digitalisieren, sondern auch komprimieren und dekomprimieren kann.
MOS (Mean Opinion Score): Bewertungsverfahren auf Basis der subjektiven Urteile von Testpersonen.
Packet switching: Ein Paket-basierter Datendienst ist eine Kommunikationstechnik, die keine feste End-zu-End-Verbindung zwischen Sender und Empfänger braucht. Stattdessen werden Datenpakete in das Übertragungssystem gebracht und darin anhand einer Adressinformation im Kopf des Paketes von Zwischenstation zu Zwischenstation weitergeleitet, genauso, wie Briefe anhand der Adresse auf dem Umschlag von Postamt zu Postamt weitergegeben werden.
PAMS (Perceptual Analysis Measurement System): Wahrnehmungsanalysesystem der British Telecom.
PSQM (Perceptual Speech Quality Measurement): Wahrnehmungsorientierte Messung der Sprachqualität laut KPN Research.

Echo kann elektrischen oder akustischen Ursprungs sein. Elektrische Echos entstehen oft durch schlechte Impedanzanpassung oder Übersprechen. Akustische Echos können durch akustische Rückkopplung beim Empfänger entstehen, beispielsweise in einem Freisprechtelefon. Echo stört umso stärker, je lauter es ist und je später es eintrifft. Ein lokales Echo ist deswegen kein Problem: seine Verzögerung ist so kurz, dass es schon sehr laut sein müsste, um zu stören. Paketbasierte Sprachnetze mit ihren Signallaufzeiten, die typisch zehnfach länger sind als bei herkömmlichen Netzen, sind in diesem Zusammenhang problematischer.

Erwünschte und störende Echoeffekte

Dabei ist es im Grunde einfach, mit Echos umzugehen, zumindest im Grundsatz: Man kann sie unterdrücken oder kompensieren. Technisch ist Echounterdrückung das einfachere Verfahren; hierbei wird der Empfangskanal abgeschaltet, während der Sendekanal aktiv ist. Das Kritische an dieser Vorgehensweise ist, dass die Schaltung zur Echounterdrückung eine merkliche Zeit braucht, um festzustellen,

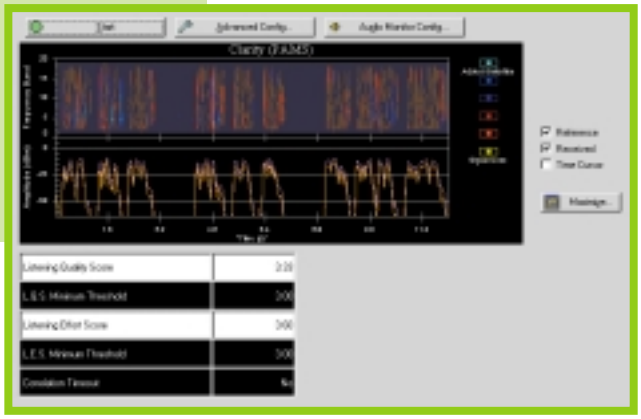
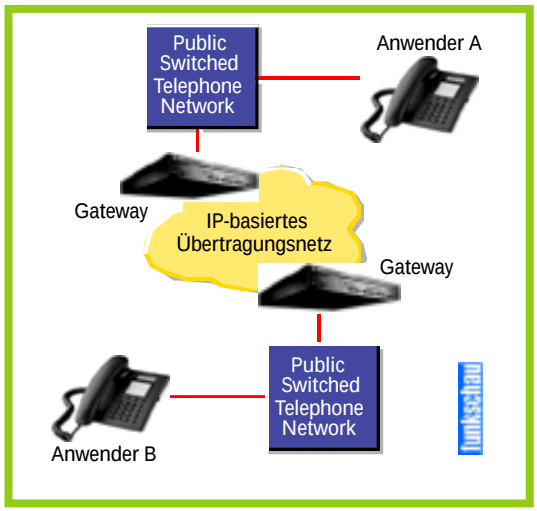


Bild: Agilent

Das Perceptual Analysis Measurement System liefert einen Wert für die Hörqualität und die Höranstrengung



Quelle: Agilent

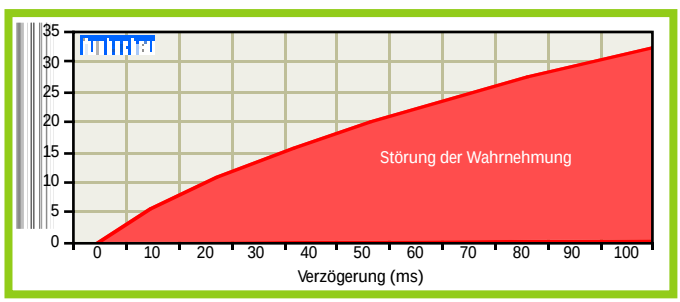
Echos, die bei Anwender A Störungen verursachen, entstehen oftmals in der Ortsvermittlung von Anwender B

dass der Sender aufgehört hat zu sprechen. Aus dieser Verzögerung kann eine Halb-Duplex-Kommunikation – Einwegverbindung in wechselnden Richtungen – entstehen, und die ist nicht so wesentlich anders als eine Verbindung mit hoher Signallaufzeit.

Ein besseres Verfahren ist die Echokompensation. Dabei wird das Sendesignal zwischengespeichert und dann vom zurückkommenden Echo subtrahiert, das schwächer und verzögert auf den Empfangskanal eintrifft. Weil leistungsfähige digitale Signalprozessoren billig genug geworden sind, hat sich dieses Verfahren gegenüber der Echounterdrückung durchgesetzt. Echokompensation funktioniert am besten bei kurzen Echolaufzeiten. Es beseitigt sozusagen die Resteffekte, nachdem andere Techniken zur Reduktion von Laufzeiten ihr Möglichstes getan haben. Die Schaltungen sind unterschiedlich leistungsfähig, beispielsweise unterscheiden sie sich in der Zeit, die sie das Sendesignal vorhalten können. Entsprechend sind individuelle Implementationen unterschiedlich geeignet zum Einsatz in einem vorgegebenen Paket-basierten Sprachnetz.

Sprachverständlichkeit ist schwer zu qualifizieren

Von den drei Faktoren, die Sprachqualität ausmachen, ist die Verständlichkeit der am schwierigsten objektiv zu fassende Parameter. Verständlichkeit besagt, wie gut der Empfänger die Wörter versteht, den Sprecher erkennt oder Feinheiten wie beispielsweise dessen momentane Stimmung erfassen kann. Bei der Verständlichkeit ist insbesonders die stark nichtlineare Beziehung zwischen Ursache und Wirkung beachten. Viele Digitaltechniken, die mit Datenkompression arbeiten , speziell MPEG-2, zeigen einen so genannten Schwelleneffekt. Mit zunehmender Signalverschlechterung sinkt die Verständlichkeit zunächst nur langsam, über einen bestimmten Punkt hinaus aber wird die Sprache auf der Empfängerseite schnell unverständlich. Die genaue Lage des „Umschlagpunkts“ wird in der Regel experimentell bestimmt.



Quelle: Agilent

Wie stark ein Echo stört, hängt von seiner Verzögerung ab und wie laut es im Vergleich zum Originalsignal ist. Die rote Linie ist die Grenze des Annehmbaren

In der Vergangenheit hat man die Verständlichkeit mit einer Technik bestimmt, die man „durchschnittlicher Meinungswert“ nannte – Mean Opinion Score (MOS). Hierfür hat eine Gruppe Testhörer ein Sprachmuster auf einer Skala von 1 bis 5 bewertet – 1: sehr schlecht, 5: ausgezeichnet, 4 entspricht der Tonqualität wie im normalen Telefonnetz. Nachteil des Verfahrens: Ein guter MOS-Test erfordert eine Menge Aufwand.

Um diese Unzulänglichkeiten zu umgehen, hat man verschiedene Computer-basierte Methoden entwickelt, mit denen man objektiv und reproduzierbar die empfangene Sprachqualität messen kann. Die meisten davon definieren einen akustischen und wahrnehmungsorientierten (kognitiven) Prozess für menschliche Sprache, um zu bestimmen, wie gut ein empfangenes Sprachmuster für einen menschlichen Zuhörer mit dem Original übereinstimmt. Zwei Messverfahren für Sprachverständlichkeit werden momentan in größerem Umfang eingesetzt. Das erste Verfahren, Perceptual Speech Quality Measurement (PSQM), wurde in den Niederlanden von KPN Research entwickelt. Mittlerweile ist es spezifiziert in ITU-T P.861. Die zweite Methode heißt Perceptual Analysis/Measurement System (PAMS), entwickelt von British Telecom in Großbritannien. Beide Verfahren arbeiten mit natürlicher Sprache oder sprachähnlichen Mustern als Eingangssignalen.

In der Praxis wird ein vorgegebenes Sprachmuster auf den Übertragungsweg geschickt, wo es durch Codierung, Paketierung, Übertragung und Decodierung verschiedene Beeinträchtigungen erfährt. Die Algorithmen zur Berechnung der Verständlichkeit werden dann mit dem Originalsignal und dem empfangenen Signal gefüttert. Ein typischer Test besteht aus Sprachmustern von männlichen und weiblichen Sprechern, wobei die einzelnen Äußerungen sorgfältig so ausgewählt sind, dass möglichst alle phonologischen Muster vorkommen.

Algorithmen für die Qualitätsbewertung

Bei beiden Methoden ist der erste Schritt, die Zeitachsen von Originalsignal und Empfangssignal anzugleichen. Das ist nicht so einfach, wie es zunächst scheint, denn der Algorithmus weiß zunächst nicht, wie viel Verzögerung im System aufgetreten ist. In Paket-basierten Netzen kommt hinzu, dass die einzelnen Pakete unterschiedlich lang laufen können, daher kann das empfangene Signal im Vergleich zum Original unterschiedlich lang werden. Man muss also die Zeitachsen solange gegeneinander verschieben, bis die Kreuzkorrelation – ein mathematisches Maß für Ähnlichkeit – ihr Maximum erreicht.

Im zweiten Schritt werden die Amplituden von Original und empfangenem Signal aneinander angepasst. Auch das ist

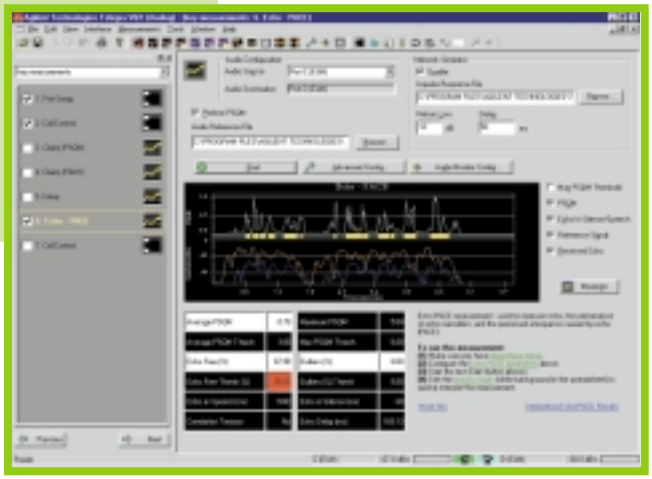


Bild: Agilent

Die Perceptual-Speech-Quality-Messung liefert einen Verständlichkeitsindex (obere Kurve), der beispielsweise umso schlechter ausfällt, je stärker Echoeffekte auftreten

schwieriger, als es sich anhört, weil das Empfangssignal in aller Regel Beeinträchtigungen erlitten hat und somit nicht einfach als gedämpfte Kopie des Originals angesehen werden kann. Trotzdem kann man die Signale einigermaßen zuverlässig auf die gleiche Amplitude bringen.

Danach werden die zeitbasierten Signale fouriertransformiert und die entstehenden Spektren in unterschiedlich breite Frequenzbänder eingeteilt, gemäß der so genannten „Bark-Skala“. Bekanntlich kann das menschliche Gehör tiefe Frequenzen besser auflösen als hohe, also sind die Bänder im Tieftonbereich schmaler, die im Hochtonbereich breiter.

Im letzten Verarbeitungsschritt kommt der wesentliche Teil der Analyse: Die Inhalte der einzelnen Bereiche werden mit Hilfe eines Sinnesmodells verglichen und verarbeitet, und damit bestimmt, wie wesentlich die Unterschiede für ein menschliches Ohr sind. Dieser Verarbeitungsschritt liefert als Ergebnis einen numerischen Messwert für die Verständlichkeit für jeden Teil der Sprachmuster.

Der PSQM-Algorithmus liefert einen numerischen Wert im Bereich von 0 bis 6,5; kleinere Werte zeigen eine bessere Sprachqualität an. Ursprünglich war PSQM dafür gedacht, verschiedene Codecs zu bewerten und zu vergleichen, nicht ganze Telefonnetze. Um auch Netze testen zu können, hat man das Verfahren daher an verschiedenen Stellen zu PSQM+ erweitert. Die Beziehung zwischen PSQM und dem traditionellen MOS ist nicht linear, man kann entsprechende Werte also nicht direkt vergleichen.

PAMS liefert zweierlei Werte: einen Wert für die Hörqualität – Ylq: listening quality – und einen Wert für die Höranstrengung – Yle: listening effort. Beide Werte liegen im Bereich von 1 bis 5, höhere Werte bedeuten bessere Qualität. Ähnlich dem Verständlichkeitsindex PSQM misst der Hörqualitätsindex von PAMS, wie gut das empfangene Sprachsignal mit dem Original übereinstimmt – wiederum nach dem Urteil eines menschlichen Hörers. Der Höranstrengungsindex dagegen zielt auf etwas anderes: Er bewertet nicht die Sprachqualität, sondern die geistige Anstrengung, die der Hörer aufwenden muss, um die Sätze zu verstehen. Dieser Wert ist besonders aussagekräftig, wenn es um sehr schlechte Signale geht.

Der Wert objektiver Bewertungsmethoden für Sprachqualität wie PSQM oder PAMS bemisst sich daran, wie genau ihre Ergebnisse mit Messwerten von gut durchgeführten subjektiven Testmethoden übereinstimmen, wie etwa dem MOS-Test mit menschlichen Testhörern. Dabei ist Erfreuliches zu vermelden: Diese wahrnehmungsorientierten Algorithmen zeigen typischerweise eine sehr hohe Übereinstimmung (größer 0,9) mit subjektiven Messmethoden. Herkömmliche objektive Bewertungsmethoden hingegen – wie die Messung des Rauschabstands – stimmen mit subjektiven Bewertungen eher schlecht überein.

Fazit und Ausblick

Die Entwickler von PSQM und PAMS, KPN Research und British Telecom, arbeiten gemeinsam an einem neuen ITU-T-Standard für ein objektives Bewertungsmodell für Sprachqualität: Perceptual Evaluation of Speech Quality (PESQ) – wahrnehmungsorientierte Bewertung der Sprachqualität. Dieses Modell verbindet die besten Teile seiner Vorgänger, darunter das Wahrnehmungsmodell von PSQM und die Routine zum Angleichen der Zeitachsen von PAMS. PESQ wird daher vermutlich Werte messen, die noch besser mit subjektiv ermittelten Messwerten übereinstimmen. *(MK)*

Codierung und Kompression von Sprachdaten

Im stationären Telefonnetz wird seit über 20 Jahren digitale Codierung von Sprachdaten und Datenkompression genutzt. Mit Ausnahme des Teilnehmeranschlusses werden praktisch alle Telefoniesignale digital übertragen. Bei der Standardcodiermethode in herkömmlichen Telefonnetzen werden die Sprachdaten mit 8 kHz Samplingrate abgetastet und in 8-Bit-Worte umgesetzt. Die Methode ist im Detail dargestellt im ITU-T-Standard G.711 (International Telecommunication Union). Aus dieser Umsetzung resultiert der bekannte Datenstrom von 64 kBit/s, den man in der Telefonie „DS 0“ nennt – die kleinste Datenrate in der Hierarchie der Digitalsignale.
Oftmals wird auch mit dem Begriff „Sprachkompression“ digitale Signalverarbeitung gemeint, dabei hat man solche Verfahren schon in der Analog-Ära der Telefonnetze eingesetzt. Das menschliche Ohr hört Frequenzen bis 20 kHz, in Telefonnetzen aber ist der Frequenzgang per Tiefpass auf 4 kHz begrenzt. Das vereinfacht maßgeblich die Konstruktion von rauscharmen Verstärkern und Schaltungen zur Frequenzgangglättung und senkt die Kosten, ist allerdings mit einer gewissen Beeinträchtiung der Sprachqualität verbunden. Zwar liegt der größte Teil der Schallenergie der menschlichen Stimme unterhalb von 4 kHz, doch der Rest höherer Frequenzen beeinflusst doch nennenswert die Verständlichkeit.

Paket-basierte Netze komprimieren Sprachdaten normalerweise über die Tiefpassfilterung hinaus weiter. Die dabei verwendeten komplexen Algorithmen erhalten nur die Teile des Eingangssignals, die für einen menschlichen Zuhörer wichtig sind, beziehungsweise gewichten sie stärker. Ziel ist nicht, Signalformen und Spektren des Eingangssignals originalgetreu wiederzugeben, sondern nur gerade so viele Daten zu übertragen, wie für eine zufriedenstellende Rekonstruktion des Signals erforderlich sind. Kompressionsalgorithmen zu erstellen, ist damit nicht nur eine Frage der Elektronik, sondern auch der Psychoakustik.
Der Zweck der Kompression ist – wie früher auch – Bandbreitenersparnis: Es sind Algorithmen im Einsatz, die Sprachdaten auf bis zu 8 kBit/s komprimieren, und das bei einer Sprachqualität, die für viele Anwendungen ausreicht; das ist eine erhebliche Verringerung im Vergleich zu den 64 kBit/s, die eine Digitalisierung nach G.711 erzeugt.
Dennoch hat Datenkompression natürlich ihren Preis: Normalerweise sinkt die Sprachqualität mit der Verringerung der Datenrate. Weiterhin dauert das Codieren für eine vernünftige Sprachqualität bei höheren Kompressionsraten länger, was natürlich die Laufzeit verlängert, und auch mehr Rechenkapazität auf beiden Seiten braucht, die dann wiederum Geld kostet.