White Paper
Voice Quality

Sound design is an art form at Snom and is at the core of our development utilising some of the world's most advance voice quality engineering tools

White Paper - Audio Quality

Table of Contents

What do we mean by Voice Quality

Voice Quality is not easily defined and whether voice quality is good or not can be highly subjective, and yet every one of us knows when we use the phone and we don't have good voice quality! When face to face the sound of someone's voice reaches us after traveling through the speaker's vocal tract. In telephony we consider good voice quality as having the ability to send and received across the telephony network without distortion or loss of the characteristics that define an individual's voice when face to face. In Telecommunications the sound of someone's voice reaches us after travelling not only through the vocal tract but also through the telephone, network and speakers, so these effect the voice quality and any changes made by them to the acoustic characteristics of speech become part of the voice quality we hear when on the telephone.

What Quality is Good Enough

Ideally the telecommunications networks should not affect the voice quality of the speaker, to maintain voice quality, therefore we must look to the network, the telephone and the speakers to maintain voice quality. But there is another consideration: When we introduced data compression into networks to accommodate digital transmission and to allow more calls to be sent over less bandwidth we introduced another component that affects voice quality - and the term Toll Quality came into use. A toll quality call is a voice call having quality comparable to that of an ordinary long distance call, originally placed over the analogue circuit-switched public switched telephone network (PSTN) where there was no compression or digitisation.

The aim of every VoIP Vendor, and the claim of many vendors is to provide Toll Quality Voice. But it is not subjectively measurable. A common benchmark telephony vendors and carriers use and the ITU has adopted to determine the quality of is the mean opinion score (MOS). MOS is a test that has been used for decades in telephony networks to obtain the human user's view of the quality of the network. A MOS score of 4 is perceptible but not annoying and 5 is rated as excellent. But MOS provides a subjective measurement based on a single set of circumstances. For instance the MOS score given in a quite office and that given in an office with extensive background noise would be different. The ITU-T standardized the MOS evaluation process in the Conversation Opinion Test, which is documented in Annex A of Recommendation P.800. A companion, and more objective, standard is P.862, the Perceptual Evaluation of Speech Quality (PESQ), which addresses the effects of filters, jitter, and coding distortions.

Measuring Voice Quality with Voice over IP (VoIP) is more objective than it could be with analogue solutions and we can use a calculation based on performance of the IP network over which it is carried. Like most standards, implementation is somewhat open to interpretation by the manufacturers. Even more significant, depending on the implementation by the IP Phone manufacturers, a calculated MOS of 3.9 in a VoIP network may actually sound better than the formerly subjective score of > 4.0 that was considered to be the equivalent to Toll Quality.

Mean opinion score (MOS)		
MOS	**Quality**	**Impairment**
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying

1	Bad	Very annoying

The output from comparative telephone voice quality measurement can be seen in the chart below. Later in this paper we will discuss the TOSCA testing that is done to determine the MOS measurement in telephony.

ETSI gives qualitative values based on TOSQA			
Handsfree WB	**Yealink-T19**	**Grandstream 1405**	**snom D715**
send	3.2	3.2	3.4
receive	2.9	3.1	3.2

What can effect Voice Quality

By looking at what voice quality is we have been able to identify the items that can effect voice quality. These include the compression used, which is determined by the CODEC, the acoustic effects of the telephone design, the impact the network itself and the acoustic changes introduced by the speakers used to output the voice. There is also a further consideration, since the brain receives and processes all sound, and only then processes out what it identifies as not being part of the conversation, the background noise both at the sending and at the receiving end of the call will impact on voice quality achieved.

Building the Telephone

The design of the telephone has a significant effect on audio quality, this includes aspects such as the thickness of the plastic selected and the shape of the phone. For best quality IP Phone design an audio engineer is involved with the industrial designer from the first stage of each new phone design. The audio engineer can explain the audio rules to the designer. For instance every speaker needs a chamber to create depth of voice, the curves on the phones will affect how audio signal is reflect, and the thickness of the plastic used is critical to the final distortion measurement. Telephone design is a trade-off between the rules of audio and the aesthetic vision of the designer. It is this seeking for high quality audio combined with pleasing aesthetic design that forces IP Phone developers to improve and come up with new solutions that take them beyond today's knowledge on achieving high quality audio. A good quality telephone design process will involve testing of voice quality in an audio lab at each stage of the telephone development to understand and optimise the impact the design has on the voice quality provided by the telephone.

Selecting the CODECs

The word codec is a shortening of 'compressor-decompressor' or, more commonly, 'coder-decoder'. A codec encodes a data stream or signal for transmission, storage or encryption, and then decodes it for playback or editing.

As with conventional telephony, with VoIP the speech is initially captured in analogue form with a microphone. This analogue information is then transferred into a digital format by a converter and changed through codecs into corresponding formats. Depending on the codec used, the data can be compressed to differing extents in this process. Most codecs use a procedure through which

information not important for the human ear is omitted. This reduces the amount of data and thus reduces the bandwidth required for transfer. However, if too much information is omitted, the speech quality will suffer. Different codec procedures handle the audio compression with different levels of efficiency. Some are specifically designed to achieve a low bandwidth at any cost. Depending on the codec, therefore, the bandwidth needed and the speech quality will vary. Although it might seem logical from a financial standpoint to convert all calls to low-bit rate codecs to save on infrastructure costs signal distortion and loss of voice quality can quickly offset this advantage The design skills of the IP Phone manufacturer in the management of codecs creates a clear differentiation between vendors. The design skills of the network engineer in handling delays in the network, including those created by the use of codecs, provides a clear differentiator between system integrators.

Mean Opinion Scores (for one implementation of different codecs)			
Compression Method	**Bit Rate (kbps)**	**MOS Score**	**Compression Delay (ms)**
G.711 PCM	64	4.1	0.75
G.726 ADPCM	32	3.85	1
G.728 LD-CELP	16	3.61	3 to 5
G.729 CS-ACELP	8	3.92	10
G.729 x 2 Encodings	8	3.27	10
G.729 x 3 Encodings	8	2.68	10
G.729a CS-ACELP	8	3.7	10
G.723.1 MP-MLQ	6.3	3.9	30
G.723.1 ACELP	5.3	3.65	30

Managing Different Onsite Conditions

Methods such as jitter buffers, echo suppression, echo cancellation, handling of silence suppression and packet loss concealment can be used in IP telephone design to improve voice quality and handle the changing performance characteristics and background interference of the site at which the telephone is used.

All the following voice techniques enhance and improve voice quality, and are quantifiable and measurable components of high quality IP Phone design and should be viewed as absolute requirements in professional and enterprise telephones.

Jitter Buffers and Packet Loss Concealment

Transmitting high quality voice over IP is made more difficult due to packet loss and jitter. Packet loss occurs when one or more packets of data travelling across a computer network fail to reach their destination. Jitter is the variation in the delay of received packets. The sending side transmits packets in a continuous stream and spaces them evenly apart. Because of network congestion, improper queuing, or configuration errors, the delay between packets can vary instead of remaining constant.

A technique used to reduce jitter involves buffering audio packets at the receiving telephone, so that slower packets arrive in time to be played out in the correct sequence at the appropriate times. The objective of jitter buffering is to keep the packet loss rate low and so improve the voice quality. A fixed method, which uses a fixed buffer size, is easier to implement than an adaptive method, but will result in less satisfactory audio quality because there is no optimal delay when network conditions vary with time. Snom telephones support adaptive jitter buffers which although more complex and

expensive to implement perform continuous estimation of the network delays and dynamically adjust the playout delay at the beginning of each transmission so ensuring a high quality of voice.

Packet loss concealment (PLC) is a technique to mask the effects of packet loss in VoIP communications. Packet loss in IP Telephony will typically have a slowly degrading impact on speech communications. The human ear is very good at handling the short gaps that are typical of packet loss. So it may take a significant amount of packet loss for the user community to be annoyed enough to report it. Packet loss will typically be experienced as clicks and glitches in the conversation. Because the voice signal is sent as packets on a VoIP network, they may travel different routes to get to destination. At the receiver a packet might arrive very late, corrupted or simply might not arrive. This could also happen where a packet is rejected by a server which has a full buffer and cannot accept any more data. In a VoIP connection, the receiver should be able to cope with packet loss.

Echo Suppression and Cancellation

Echo suppressors work by detecting a voice signal going in one direction on a circuit, and then inserting loss in the other direction. This added loss prevents the speaker from hearing his own voice. Echo cancellation is based on recognizing the originally transmitted signal that re-appears, with some delay, in the transmitted or received signal. Once the echo is recognized, it can be removed by subtracting it from the transmitted or received signal. IP Phones echo controls are implemented digitally using a digital signal processor (DSP) or software and at Snom we implement to the ITU requirements. Digital signal processing is the mathematical manipulation of the information signal to modify or improve it. DSP is not one size fits all. Different DSP coefficient pre-sets are needed for different room types. Refining the voice using these techniques will improve the subjective quality, as an additional benefit the process also increases the effective use of bandwidth as silence suppression prevents echo from traveling across the voice network. By preventing echo from being created or removing echo if it is already present voice quality is improved.

Automatic Noise Reduction

Voice quality can be further enhanced by reducing the surrounding noise coming from the other parties' telephone. Automatic Noise Reduction (ANR) is the technique used to reduce this hiss noise (white noise) caused by the phone's electronics, coding/encoding and background noise, unwanted audio picked up by the microphone. This noise can be reduced using digital signal processing (DSP). ANR requires fine tuning, a good noise reduction algorithm doesn't change the speech, it only removes the unwanted noise, but aggressive noise reduction will introduce artifacts and distort speech, reducing the voice quality. The ANR needs to differentiate between the hiss noise introduce by the telephone hardware and de/encoding and background noise, since for each noise type different algorithms are needed. Static noise reduction algorithms are best for hiss noise reduction and dynamic algorithms providing adaptive noise reduction and dynamic range compression for background noise reduction.

Silent Suppression

Most conversations include about 50% silence. Silence suppression is an application used in telephony to prevent transmission of 'silent' packets over the network when one of the parties involved in a telephone call is not speaking, thereby reducing bandwidth usage. When silence suppression is on, comfort noise needs to be generated locally by the IP Telephone at the other end of the call so that the other party will not mistakenly believe that the call has been terminated.

Synchronisation

By its very nature VoIP is an asynchronous connection, and sometimes audio packets are dropped simply because the sender and receiver are not using the same clock. This can be resolved by good use of packet loss concealment and jitter buffer management.

Network Requirements for Voice Quality

Bandwidth in the IP world is largely shared, so congestion and delay are often present and can cause problems for multimedia applications such as voice. Voice traffic is more sensitive to delay and jitter, but can tolerate loss, whereas even a small loss of data can corrupt an entire file or application. This means that the characteristics required for a high performing data network and those for a high voice quality network are not the same.

The following chart illustrates the impact of the network on Voice Quality:

ETSI examines speech quality in presence of network impairments (packet loss, jitter) with TOSQA values						
Test Scenarios	**Jitter**	**Packet loss**	**Ye-T19**	**Gr-str 1405**	**Snom D715**	**ETSI 202-739 Requirements**
Clean conditions	--	--	3,4	3,5	4,4	>3,6
20ms jitter 0% PL	20ms	--	3,4	3,2	4,4	>3,4
0ms jitter 1% PL	--	1%	3,4	3,5	4,0	>3,4
20ms jitter 1% PL	20ms	1%	3,3	3,2	3,9	>3,4
20ms jitter 3% PL	20ms	3%	3,2	3,1	3,6	>3,2

Sufficient Bandwidth

A first requirement is to provision sufficient network bandwidth to support real-time voice traffic. For example, an 80-kbps G.711 VoIP call (64 kbps payload plus 16-kbps header) will be poor over a 64-kbps link because at least 16 kbps of the packets (which is 20 percent) will be dropped. This example also assumes that no other traffic is flowing over the link. After you provision sufficient bandwidth for voice traffic, you can take further steps to guarantee that voice packets have a certain percentage of the total bandwidth and get priority.

Prioritizing Voice Traffic

Prioritizing voice traffic ensures that jitter and packet loss is reduced. In order to prioritise the voice traffic it needs to be identified, this identification process is called *packet classification*. After a packet has been classified, the packet can be marked by setting designated bits in the IP header. I ensure voice quality this identification needs to be done as far out towards the edge of the network as possible.

Quality of Service Mechanisms

Quality of service (QoS) is the overall performance of a telephony or computer network, particularly the performance seen by the users of the network. Networking vendors' *QoS* products manage the delay, jitter, bandwidth, and packet loss parameters on the network. QoS involves giving preferential treatment through queuing, bandwidth reservation, or other methods based on attributes of the packet. A service quality is then negotiated. Examples of QoS are CBWFQ (Class Based Weighted Fair Queuing), RSVP (RESERVATION Protocol - RFC 2205), MPLS, (Multi Protocol Label Switching - RFC 1117 and others).

Using VLANs

Computer networks can be segmented into local area networks (LAN) and wide area networks (WAN). Network devices such as switches, hubs, bridges, workstations and servers connected to each other in the same network at a specific location are generally known as LANs. An LAN is also considered a broadcast domain. A VLAN allows a network of computers and users to communicate in a simulated environment as if they exist in a single LAN and are sharing a single broadcast and multicast domain. A VLAN allows several networks to work virtually as an LAN. One of the most beneficial elements of a VLAN is that it removes latency in the network, which saves network resources and increases network efficiency and therefore reduces delay and jitter and improves voice quality. In addition, VLANs are created to provide segmentation and assist in issues like security, network management and scalability. Traffic patterns can be controlled by using VLANs, creating a separate VLAN for voice reduces the amount of broadcast that the telephone will receive. A voice VLAN can be manually applied to an IP telephone or provided by a DHCP server.

Building the Telephone

At Snom evaluating voice quality on a new product begins as soon as a first injection of plastic is produced and continue throughout the life cycle of the product. We are aware of the complexity of VoIP terminal devices and consider it a critical aspect of our telephone design. We have improved audio quality over the years by combining our acoustic experience with the latest DSP algorithms and our VoIP signalling know-how. Specifically, we have solved various issues inherent in VoIP technology, including processing delay, network delay, network packet loss, need for VAD and CNG, and countless types of noise. And, of course, we have addressed the main issue of synchronization, as by its very nature VoIP is an asynchronous connection, and sometimes audio packets are dropped simply because the sender and receiver are not using the same clock.

It is the attention to detail during design that creates voice quality differentiation between telephones. The Snom7xx, for example, has been built to pass the frequency response requirement based on ETSI 202-379 at every telephone-to-ear pressure. The telephone uses a specially designed high leak receiver that allows for the best sound quality at each telephone. We use the most realistic artificial ear type during tests, too, which makes the receive curve extremely difficult to surpass.

A further consideration in telephone design is avoiding stability loss. Most frequently noticed when the handset is laid on a flat surface, such as a desktop, and produces positive feedback between the microphone and the speaker. Stability loss is a measure of the contribution of the telephone set to the overall network stability requirements. Stability loss is defined as the minimum loss from the digital input (receive) to the digital output (send) at any test frequency.

On another front, high quality jitter buffer and packet loss concealment software in the Snom 700 Series have been improved to operate in poor network conditions and the Snom speakerphone has excellent double talk performance, and algorithms such as background noise cancellation and adaptive gain control provide for voice clarity in every condition.

Subjective tests are as important as objective tests, and a good objectively-tuned phone can still provide bad audio. At Snom, well-tuned audio devices mean a cycle of objective tuning followed by subjective sessions, until the job has is finished.

Using an Audio Lab to Test Voice Quality

VoIP Audio Measurement equipment has evolved in sync with VoIP technology, today's audio labs require extensive investment in industry leading software and equipment. For modern telecommunications, old audio standards such as TIA-810B (Narrow band) and TIA-920 (Wide Band) fail to match modern business expectations. These standards are focused on half duplex connections. Important aspects of the audio quality are not exposed, and many typical problems remain unresolved. TIA-based audio optimized devices are unable to match customer expectations for high quality audio. Today's audio labs test for wide-band audio based as defined by ETSI 202 739 and ETSI 202 740.

With ETSI, all the original requirements from the TIA standard are covered, plus ETSI extends the requirements in frequency response domain and in loudness ratings, which requires high quality electroacoustic converters. ETSI also includes double talk behaviour measurements and speech quality in presence of network impairments (packet loss, jitter) and, at the end, speech quality in presence of the background noise.

Snom have invested in tools to measure and design end points that provide no-compromise audio quality. Today the Snom audio lab uses software and equipment to measure and design our telephones and to ensure we create devices with no-compromise audio quality. In addition we can provide spectral echo attenuation over time testing and modify our early telephone designs to fix all over-limits distortions. This investment in best in the market tools allows Snom to design for high audio quality from the very earliest phase of the IP telephone development.

Sound and Vibration Analysis

The software Snom uses for acoustic testing and tuning enables sound and vibration analysis (data acquisition, analysis, playback, reporting, data management), and enables hearing and analysing at the same time, this is important since the acoustic impression helps to reliably identify sound problems or to define target sounds. The tests implemented cover all conversational speech quality aspects such as delay measurements in sending and receiving direction, one-way speech quality tests under single talk conditions in sending and receiving direction, echo tests, quality during double talk, and quality of background noise transmission. High quality jitter buffer and packet loss concealment software in the Snom 7 Series have been enhanced using our state of the art network simulator to operate effectively even in very poor network conditions.

Objective Testing of the Network for Voice Quality

Snom also have the ability in-house to test to the Telecommunications Objective Speech Quality Assessment (TOSQA) standard. TOSQA enables us to objectively measure the Mean Opinion Score (MOS) mentioned earlier in this paper. TOSQA is an objective speech quality measurement method that can be applied to scenarios where a speech signal is transmitted via a measurement object and recorded at an electrical interface; for example over the network.

Generally speaking, the TOSQA method is based on a speech signal which is fed in as measurement signal at one point of a transmission path, transmitted via the path (e.g. the network) and recorded at another point as a transmitted signal. The recorded signal and the fed-in signal (also called the reference signal) are compared and from this comparison the quality values (TOSQA value) is calculated. If the recorded signal shows a good correlation with the fed-in signal (i.e. it was only slightly “falsified”, distorted or otherwise impaired during the transmission), high TOSQA values are achieved. If the signal was severely impaired by the transmission path, i.e. strongly modified compared to the fed-in signal, the comparison of the transmitted signal to the fed-in signal obviously shows strong differences and the algorithm calculates a loss of quality, represented by a lower TOSQA value.

Testing for End to End Voice Quality

At Snom we also test to PESQ as defined in ITU-T Recommendation P.862. PESQ stands for 'Perceptual Evaluation of Speech Quality' and is an enhanced perceptual quality measurement for voice quality in telecommunications. PESQ was specifically developed to be applicable to end-to-end voice quality testing under real network conditions, like VoIP. PESQ was particularly developed to model subjective tests commonly used in telecommunications (e.g. ITU-T P.800) to assess the voice quality by human beings. Consequently, PESQ employs true voice samples as test signals. In order to characterize the listening quality as perceived by users, it is of paramount importance to load modern telecom equipment with speech-like signals. PESQ was particularly developed to model subjective tests commonly used in telecommunications (e.g. ITU-T P.800) to assess the voice quality by human

beings. PESQ employs true voice samples as test signals to characterize the listening quality as perceived by users.

Snom utilises all these tools and software in the design of our IP Telephones and this ensures we deliver the best quality to our customers according to the latest requirements of modern telecommunication.

Snom Technology AG
Wittestr. 30 G | 13509 Berlin | Phone +49 30 398 33-0 | Fax +49 30 398 33-111
office.de@snom.com