YAMAHA

WHITE PAPER

AUDIO QUALITY MATTERS

The need for quality audio and voice over distance continues to grow

Recording and transmitting audio, primarily for human-to-human interactions, has been a technology available for 120 years. However, from the beginning, the process of moving audio from one location to another has involved compromise. Whether limited by the technologies of the time or, more often, by the need to maintain backward capability to a large deployed user base, technology has moved forward at a relative snail's pace.

While audio and voice transmission has been limited in adoption of new technologies, the need for quality audio and voice over distance continues to grow. Research has shown that voice is the mechanism humans use to communicate information for task-based activities. In 1974, as part of a study to see how humans would best communicate with a machine, Robert Ochsman and Alfonso Chapanis conducted an experiment to explore which communications mechanisms would best enable a sender of information to communicate with a seeker who needed the information to complete a task. The experiment placed the sender and seeker in separate rooms and the communication interface was changed over many repeated experiment attempts. First they used a teletype (it was 1974, after all), then a handwriting machine, then both, then voice, then voice and others, and finally they opened a curtain over a window between the rooms to simulate video.

The results of the experiment should come as no surprise. Adding voice dropped the time to completion by about 70 percent on average, but the fact that "video" did not change the timing indicated a core human characteristic: humans solve problems and complete tasks by working together and TALKING. In fact, even the old adage, "A picture is worth a thousand words" shows how speech is critical. A picture of four people, dressed in business casual attire on a beach with Diamond Head in the background says much. They are probably business people, going out on the beach in Hawaii. But that is all the picture says. Who is the third person from the right? Why are they there?

If voice and human speech are the keys to task completion and operational success, we do not seem to be good at making it work. In many conversations today, the quality of the audio makes the ability of the participants to clearly achieve the objectives of the interaction a challenge. Organizations today are dependent on telephony and voice to perform virtually all of the functions of an organization. With tele-commuting, flexible hours, distributed teams, and global operations, communications are the lifeblood of a modern organization. When there is poor audio quality, the flow of information and task completion is impacted dramatically. This is indicated by the thousands of web articles on poor mobile and VoIP voice quality. We depend on voice to get things done, and poor or reduced voice quality can have even greater impacts when workers are from different regions or even different countries and the languages used may not be the primary language for some of the participants.

Clearly, poor audio is no longer acceptable in the enterprise. While there are few studies in this area, for most of us the impact of bad voice quality is readily seen. All of us have been on that conference call where no one can understand what someone is saying; in fact, it happens regularly. Employees clearly understand that remote communication is key to working efficiently. In a recent survey, 40 percent of respondents named impromptu meetings from co-workers stopping by their workspace as a major office distraction. In fact, almost half (46 percent) said they primarily communicate with co-workers through e-mail, IM or phone to avoid the distractions that come along with face-to-face interactions (like idle chatter).2 With business strategy, globalization, remote working, and workers needing to focus, the value of the phone call is increasing, but the quality has generally lagged behind, in fact it has often dropped due to cellular and poor VoIP solutions.

SETTLING FOR "OKAY" AUDIO QUALITY

We settle for poor voice quality because, for many of us, that is what we have gotten used to. While the legacy of the phone system is marginal audio and voice quality, even that has been degraded by both the mobile cellular network as well as many VoIP implementations of the past. We have learned that we have to trade off quality for convenience and low cost. In the pursuit of the goal to talk anywhere and for as low cost as possible, we continue to degrade the quality of our communications.

However, the times are rapidly changing. New technologies are expanding the options for transmission of audio and voice communications. Networks are getting faster at an exponential rate, enabling higher bandwidths and fewer congestion issues. The wireless network is increasing at a rate that is making the bandwidth used for voice communications to be an ever-smaller portion of the overall bandwidth.

The result of all of these changes is that the dawn of a new era of audio/voice communications is upon us. However, to take advantage of this, there is one element that needs to be considered: the end devices themselves. We are now at time when remote communications can replicate in the audio and voice domain equivalent to being in the same room together. While remote communications can never exceed the capabilities of talking to each other from a few feet apart, remote voice communications can replicate that experience to 99.9 percent or more. Making audio/voice communications work better is a critical aspect of the next generation of solutions. We cannot afford to have poor audio quality impact business decisions or tasks, or to have quality impact the ability of our organizations to change to meet the business challenges of an ever-increasing global business environment.

WHY TODAY'S VOICE AUDIO IS STUCK IN THE 1950s

The telephony network is an amazing invention of the human race. The concept that almost anyone can pick up a phone and call someone on the other side of the world was truly amazing…in the 1950s. In fact, today's Public Switched Telephone Network has its roots in the 1950s. That is when the large telephone companies began the definition of the move from an analog to digital telephone system. Prior to 1950, phone calls were made over wires that supported one call per pair of wires. To make a call from California to New York, relays would be set to essentially connect a wire from California to New York. The complexity and size of the technology made it a challenge to grow, and voice quality was impacted by these long wires.

In the 1950s, the digital phone network was defined. While there is much about how the system works, the focus for this paper is on the mechanism to transmit audio/voice over that network. Remember, this is digital before the microprocessor, chips, or even general computers. This was development that was 30 years before the Compact Disc (CD).

To make the transmission of voice calls in the digital domain manageable in a world that was just moving from tubes to transistors, the quality of the audio was reduced to meet the technology challenges of the transmission. A basic digital voice signal is carried in a 64 Kilobit Per Second (kbps) channel generally referred to as G.711. The data is generated by a sample rate of 8,000 samples per second and a sample of 8 bits. While this sounds like a lot, it is a very small sample. CD recordings, introduced 30 years later, use a sample rate of 44.1 thousand samples per second (expressed as kilohertz per second or kHz) and a 16-bit sample size. Samples per second defines

> *"The telephony network is an amazing invention of the human race"*

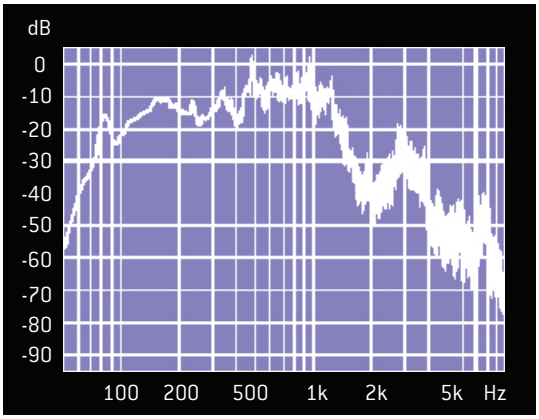
the frequency range of the transmission.

In digital encoding of an analog signal, the sample rate must be slightly higher than double the effective frequency range being converted to digital. This is known as the Nyquist ratio and says that two samples must be used to determine the state of the signal at that point. The CD uses 44.1 kHz as the sample rate with the half value of 22 kHz so that a filter to remove the noise of the sampling results in a frequency range up to about 19-20 kHz, generally considered the limit of human hearing. The range of the human voice extends from 80 Hz to 14 kHz; however, the 8 kHz sampling of the traditional digital phone limits audio frequencies to the range of 300 Hz to 3.4 kHz. This challenge is further exacerbated by the relatively low dynamic range that an 8-bit sample results in. In fact, a phone call has only 255 levels of volume, while the CD has over 64,000. For music, even the CD is no longer considered adequate, as modern music mixers generally use a 192 kHz sample rate and 24-bit samples. The reason for all of this is quality. Increasing the sample frequency increases the frequency spectrum; increasing the sample size enables a much better signal-to-noise ratio and much better clarity between intonations. The combination produces dramatically better quality. The inverse is equally true, as compression used in legacy wireless technology is even worse in both size and quality, "Can you hear me now?"

So the phone system is compromised due to the time period and technology limitations of development. The devices that have traditionally been used for telephony are defined by these limitations as well. There is little reason to have wider speaker frequency response when the actual transmission is so limited, so traditionally, manufacturers have chosen components that reflect the low frequency response. This limitation of design has been extended to include the relatively low dynamic range of the 8-bit samples. In conference devices, simple microphones and speakers have been adequate; they can deliver quality equivalent to the transport network.

To clearly understand how limited the phone system is, listen to any wideband audio content over a traditional phone. Listening to music with a full frequency spectrum and high dynamic range demonstrates how completely the experience is masked by the poor quality of the system. In fact, there is a reason that there are specialized companies selling music on hold content; it is content that is optimized to the relatively poor audio characteristics of the current telephony network.

The reasons for this are multi-fold. The low frequency limits some parts of the human voice, the resonant frequencies that make up our "sound." While early developers of telephony rightfully noted that the PRIMARY frequencies of human speech fall in the 300 to 3.4 kHz range, their antiquated testing equipment do not tell the whole story. First, the speaking voice of a typical male adult will have a fundamental frequency from 85 to 180 Hz; a typical female adult from 165 to 255 Hz. So, in a typical conversation, much of the speech may actually be below the G.711 range. Our brains can interpret the higher frequency harmonics generated to actually hear what is said, however, this makes the sound nasally and harder to understand. Similarly, the higher frequency harmonics may be dropped, resulting in a flat sound.



Human Speech Frequency Range

Figure 1.

Figure 1 shows an actual frequency spectrum of a human male speaking as recorded over a one-minute period, showing the relative level and frequency of the audio generated. Note that while the predominant frequencies are in the 100 Hz to 1 kHz range, there is significant content beyond that range. The reality is that the overall sound of the human voice is the sum of all of these frequencies. The reason for this is a factor of both sound and human anatomies. When we talk, we generate a basic frequency, but certain sounds make much higher frequencies. For example, the hard "s" sound comes with a frequency that is much higher than the typical range. Naturally these are small variants, but they are well within the normal range of human hearing and are critical to fully understanding the spoken word. In human speech, this is further exasperated by the natural resonance of the human voice box and the mouth that "tunes" the speech. In addition, having the dynamic range to make these frequencies audible versus the fundamental is critical. The result is that while the fundamental frequencies of human speech are relative low and can be defined to fall within the

legacy telephony spectrum, the full spectrum has a range of frequencies at a wide range of dynamic levels that combine for the full speaking experience. The implementations of the telephony system have left much of these extended values on the ground, resulting in a perception of telephony as flat and "lifeless."

THE NEW AUDIO PARADIGM

The movement of telephony voice from the simple 64Kbps basic telephony to a higher resolution and more dynamic audio has been caused by VoIP and the capability of using compression as well as higher bandwidth to transport the audio traffic. In VoIP, the voice can be carried as G.711; however, even the earliest VoIP solutions enabled compression. In the early days of VoIP, this was done to reduce the bandwidth. However, as bandwidths on all networks continue to increase, the percentage of traffic that is generated by voice is decreasing. Networks tend to have a 10x increase in bandwidth every 7-10 years.

Whether desktop, nomadic, or wireless, this increase in bandwidth is driven by web, applications, and predominately in recent years, video. Clearly the number of voice calls is not increasing at this rate, so the possibility of using the available bandwidth for higher quality emerged.

The first step in higher quality voice was using compression techniques to put a higher quality voice signal into the same network bandwidth. The first attempts at HD voice used the G.711 codec technology with a higher sampling rate. The G.722 codec is generally considered the baseline for HD voice; it captures and delivers sound between 30 Hz and 7,000 Hz. While the G.722 is essentially double the frequency of a G.711 codec, an HD voice call using G.722 consumes the same 64 kbps of bandwidth as the digital POTS equivalent of G.711. So moving from G.711 in a VoIP phone system to HD voice can be done without needing more bandwidth.

While this was the start of HD back in the early 90s and was supported by a range of companies, the introduction of web-based solutions has accelerated the quality that a VoIP session can carry. For example, the SILK codec used in Skype can go up to 24 kHz sample rate, yielding an effective frequency band up to over 10 kHz. It does this with a bandwidth of around 40 Kbps plus packet overhead. Figure 2 shows the definitions of different types of codecs for audio as defined by the Internet Engineering Task Force (IETF).

The latest advances are in codecs like the OPUS codec. OPUS is an Open Source codec used in web technologies like WebRTC. OPUS is designed for both fixed and variable bit rates (where the rate changes based on the content). OPUS can operate at a sample rate of up to 48 kHz, effectively delivering a codec that can cover the full range of human hearing. In fact, audio Royalty Free – Open Source Licensing Fees – Not Open Source codecs used in both Relative Quality and Bandwidth of Different Codecstelephony and video over IP have a full range of capabilities. Figure 3 shows the range of codec options available today, from very low bandwidth to full range including stereo. The MP# and AAC codecs are typically used in audio for either

Codec Comparison

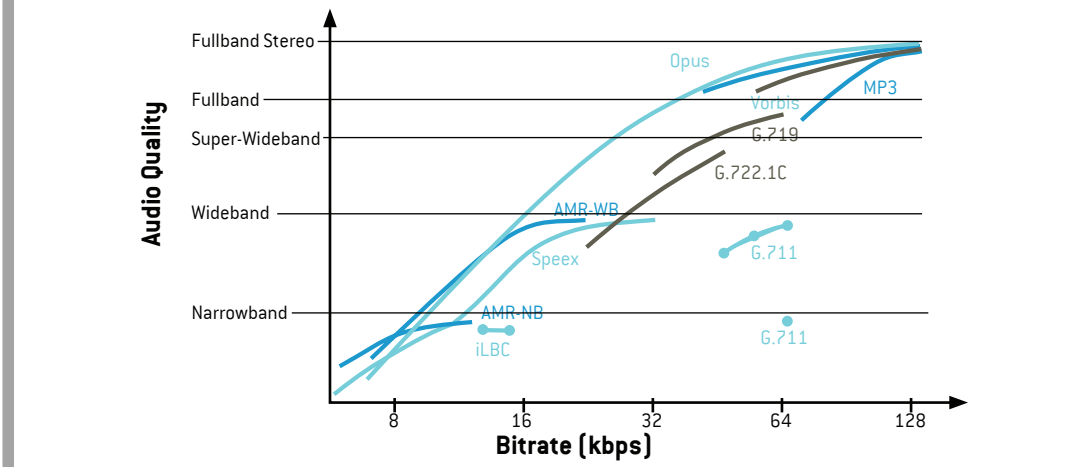
ABBREVIATION	AUDIO BANDWIDTH	EFFECTIVE SAMPLE RATE
NB (narrowband)	4 kHz	8 kHz
MB (medium-band)	6 kHz	12 kHz
WB (wideband)	8 kHz	16 kHz
SWB (super-wideband)	12 kHz	24 kHz
FB (fullband)	20 kHz	48 kHz

Figure 2.

recording or streaming of personal audio. While this chart shows frequency range, the dynamic range is also increasing, with OPUS having a selectable sample size of up to 24 bits.

The result of the movement from simple PSTN G.711 voice of the last century to the newer codecs like SILK and OPUS as well as the availability of additional bandwidth has led to many VoIP services moving to HD and beyond voice. For example, a typical Skype for Business call might use the G.722 codec in a 16 kHz sample or 7 kHz range for stereo of the Microsoft RTA codec at the same speeds for a point-to-point connection. In the emerging world of WebRTC, the OPUS codec is generally the choice and is often used in wideband or

Figure 3.



Relative Quality and Bandwidth of Different Codecs

even fullband.

The result of all of these changes, driven by inexpensive compute power, ever-increasing bandwidth and efficient low cost open source codecs is that more and more of our "telephone calls" resemble a high quality audio connection. However, as the components of the audio chain have remained unaltered in many cases, the actual audio is impacted by the devices. Thanks to digital technology, the microprocessor and software, what was once the strength of the system has now become the weak link.

UNDERSTANDING GOOD AUDIO QUALITY

For the layman, the difference between a high quality audio stream and a low quality one is often challenging, however the difference between legacy telephony audio and the modern high definition audio for IP-based collaboration is clear. In normal speech, a quality audio codec and implementation will make the human voice more understandable. This is because the system will reproduce both the full frequency range as well as the dynamic range that today's communication and collaboration systems can deliver.

For example, when a traditional G.711 call is played through a high quality speaker, the sound will be flat and understanding some speech points may be challenging. Playing the same call using an HD codec will result in a warmer and easier-to-understand conversation. However, play the same HD quality conversation through a low cost USB speaker and it will reflect the limited quality of the source. The challenge for modern communications speaker developers is to deliver a full range audio signal, something that is often achieved with multiple speakers in the high fidelity audio world. Limiting a device to a single small speaker and expecting intelligible sound is allowing the last challenge in the new audio world to persist.

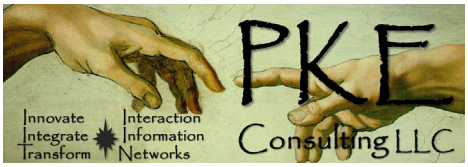
GETTING THE ENVIRONMENT RIGHT

While the audio codec, bandwidth, frequency and sampling rate define the audio signal, the environment where the user and the device exist can be more important. In a small space, while a small speaker may work, it will not have the presence to get the full nuance of speech. In a larger space the challenge is generating enough volume to fill the space without over-driving the amplification or speakers, resulting in distortion that can hide the subtleties in speech.

In small and large spaces, an external microphone can be an issue. This can show up in many ways. For example, in a small room there may be echoes in the room itself that impact hearing what is said. In a conference setting, the device used generally should have multiple microphones as well as adaptive circuitry to select not only the best microphone, but also to enable gain matching so users do not sound like they are at the end of a pipe. Managing these multiple microphones is a challenge, but also one that is amplified by the need for excellent echo cancellation.

One final key to a great audio conversation is a lack of echo. While echoes can be introduced in the legacy system, they often are generated by end points in modern VoIP systems. For example, many PCs have variable timing between the microphone and the speaker. If the application is doing echo cancellation (removing the input from the output), the large time domain and variability of the audio through the internal components can result in echoes getting back into the conversation. While

these may originate at one device, the resulting echo is heard at the other end, or on a conference call, by everyone else. In the office environment a headset is a critical tool to manage echo as the separation of microphone and speaker naturally reduces echo. However, in a conference room or using a speakerphone-type device in an office, the key to echo cancellation is managing the audio environment, something that a great conference device can do. By using multiple microphones combined with active tuning and echo cancellation, a great audio device can sound as good, or even better than a headset.



THE NEW AUDIO WORLD

In the new world of high definition audio, combined with the bandwidth and power of the Internet, a new age of voice and audio communications is exploding. The advent of the Real Time Web and multiple new applications using VoIP HD audio drive an ever-expanding need for higher quality voice audio. By combining the full frequency spectrum with higher dynamic range resolution and audio devices specifically designed for this new environment, conversations and conferences can be much easier to understand and participate in. The resolution of these new systems enables users to understand small variations and more readily identify speakers from their voice patterns, as well as understand the nuances of human speech, that if not noticed, might result in misunderstandings.

As we all go on this journey to the new advanced capabilities of a high definition voice conferencing and collaboration world, making sure that all elements of the solution are aligned to optimize the overall experience and assure understanding is critical. The selection of the device to put in a conference space, whether in a dedicated conference room or in an office will have a major impact on the collaborative events that are conducted in the space. Choosing devices with the right speaker and microphone solutions that are appropriate for all your conferencing spaces will help ensure you are getting the most value from your new communications and collaboration solution.

References:

1 Ochsman, Robert B. and Chapanis, A. (1974). “The effects of 10 communication modes on the behavior of teams during co-operative problem-solving.” International Journal of Man-Machine Studies 6: pgs. 579-619.

2 Harris Interactive conducted a nationwide survey on behalf of Ask.com in which they canvassed more than 2,060 professionals, ages 18 and up, between March 26 and March 28, 2013, to unearth the preferences and habits of U.S. office workers when it comes to an optimally productive environment.

YAMAHA

SHARING PASSION & PERFORMANCE

Yamaha Unified Communications, Inc.
144 North Road #3250
Sudbury, MA 01776 USA
+1 800-326-1088
uc-sales@music.yamaha.com

Yamaha UC EMEA
190 High Street Tonbridge, Kent
TN9 1BE, UK
+44 (0)1732 366535
uc-salesemea@music.yamaha.com

Yamaha UC APAC
+852.8108.8820
uc-salesapac@music.yamaha.com